MASTER EN SCIENCES INFORMATIQUES

Gestion des statistiques dans la conception de bases de données

Wasiak, Christophe

Award date:
1993

Awarding institution:
Universite de Namur

Facultés Universitaires Notre-Dame de la Paix
Institut d'Informatique

Rue Grandgagnage, 21
B-5000 Namur (Belgium)

Gestion des statistiques dans la conception de bases de données

Christophe WASIAK

Promoteur : J-L. Hainaut

Mémoire présenté en vue de l'obtention du diplôme de Licencié et Maître en Informatique

Année Académique 1992- 1993

Gestion des statistiques dans la conception de bases de données.

Wasiak Christophe

Résumé

Le mémoire concerne la description statistique du contenu d'une base de données existante ou en conception. Ces informations, très souvent négligées dans les méthodes de conception de bases de données, sont essentielles pour l'optimisation ou la mise au point physique par exemple, ainsi que pour l'estimation des perfomances futures. Le mémoire étend le modèle Entité-Association pour spécifier la taille des populations de types d'entités, de type d'associations, de domaines, d'attributs et de leurs relations. Nous proposons un ensemble minimal de statistiques, ainsi que les équations qui les relient. Les valeurs des statistiques sont ensuite considérées comme des informations évoluant avec le temps. Le mémoire propose aussi des équations de transformations des statistiques associées aux opérations de transformations de schémas. La gestion souple des statistiques, proposée ici, est une démarche multi-niveau : d'abord, nous fixons le sens de dérivation des équations, ensuite des valeurs valides sont introduites, ensuite, après dérivation, on obtient les valeurs des statistiques dérivées; si une valeur dérivée est non-valide, des nouvelles valeurs sont introduites et on recommence. Ce modèle statistique et la gestion souple des statistiques sont analysés pour leur intégration dans l'atelier logiciel TRAMIS.

Abstract

This paper is about the statistical description of the content of an existing or future database. This information, very often neglected in database design methods, is essential for database optimisation and physical tuning for instance, or for estimation of future performances. The paper defines an extension of the E-R model to specify the size of the populations of entity types, relationship types, domains, attributes, and their interconnections. It proposes a minimum set of statistics together with the relations between them. Data statistics are then considered as time linary-dependent quantities. The paper proposes also translation rules associated to the transformation operations on schemas. The flexible statistical management proposed is a multilevel approach : first the derivation rules are fixed one way by; next, validated values are accepted; next the derived statistics are computed; if the computed values aren't valid, new base values are asked and the process is started once again. This statistical model and the flexible strategy are then analyzed for integration in the commercial CASE tool TRAMIS.

Je tiens tout d'abord à exprimer ma profonde gratitude à mon promoteur, Monsieur Jean-Luc Hainaut, pour son travail directeur tout au long de l'année.

J'exprime également toute ma reconnaissance à Messieurs Olivier Marchand et Jean-Marc Hick, dont l'assistance a permis l'étude de la partie intégration de la gestion dans TRAMIS.

J'adresse aussi mes remerciements à Messieurs Jean-Paul Leclerc et Claude Cherton, pour leurs conseils et leur disponibilité à discuter la partie mathématique du travail.

Enfin, je remercie en particulier mes parents pour m'avoir permis de poursuivre des études.

Table des matières

Chapitre 1

Introduction

Ce chapitre nous présente le problème de la gestion des statistiques dans la conception de bases de données. La conception de bases de données a pour objectifs de représenter un système du monde réel (ou réel perçu) avec des données et de concevoir une gestion informatisée de celles-ci. Les statistiques sont les informations quantitatives sur les classes de données d'une base de données, et sont utiles pour assister le concepteur.

1. Introduction

1.1. Gestion des statistiques dans la conception de bases de données

Nous allons commencer par un rappel des étapes principales de la conception d'une base de données et poser le problème du mémoire. La description suivante a été réalisée à l'aide de plusieurs références [BOD-PIGN,83], [HAI,86] et [HAI,89].

Une **base de données** est une image ou un modèle d'un sous-système du monde réel, lequel est communément appelé le réel perçu. Le mode de représentation des connaissances statiques des bases de données est un modèle. Par exemple le modèle Entité-Association (présenté dans [BOD-PIGN,83]), peut être utilisé pour la description statique des bases de données. Une base de données est constituée par une telle description, communément appelé schéma, et par l'ensemble de valeurs du réel perçu, communément appelé extension, qui respecte à tout instant le schéma. L'extension peut évoluer très rapidement, contrairement au schéma, grâce à laquelle l'ensemble des valeurs est géré au cours du temps.

La **démarche classique de conception d'une base de donnée** consiste en trois parties consécutives, l'analyse conceptuelle qui conduit à un schéma conceptuel de la base de données (développée dans [BOD-PIGN,83], la conception logique et la conception physique dans [HAI,86]). Ces deux dernières parties constituent la phase de mise en oeuvre du Système d'Information et plus particulièrement de sa base de données.

> L'**analyse conceptuelle** conduit à élaborer une description complète du système d'information qui soit indépendante de la notion même d'outil informatique. Cette description constitue la spécification de la solution retenue. Elle est constituée d'un schéma conceptuel des données, d'un schéma conceptuel des traitements ainsi que du relevé des quantifications.
>
> La **conception logique** conduit à une solution exécutable sur une machine abstraite, strictement indépendante des machines réelles et consiste principalement en un schéma des accès nécessaires, accompagné de sa quantification statique et dynamique, d'une architecture abstraite de modules et des algorithmes effectifs des modules.
>
> La **conception physique** conduit à une solution exécutable par une machine réelle et consiste principalement en un schéma des accès nécessaires conforme à un Système de Gestion de Données, des algorithmes effectifs conformes au SGD et au langage de programmation, expression du schéma conforme dans le langage de déclaration du SGD, définition des sous-schémas destinés aux différents utilisateurs, rédaction des programmes, définition du schéma interne de la base de données (fixation des paramètres physiques).

Le **processus de conception d'une base de données** correspond à l'évolution de l'expression conceptuelle d'une solution à sa forme exécutable en machine. Ces

expressions décrivent toutes les mêmes concepts, dont la forme est liée à l'intégration de certains contraintes conceptuelles, logiques ou techniques des trois phases décrites.

Les **transformations** sont les outils de base de la conception de bases de données, qui permettent par des règles systématiques la génération d'une expression à l'autre. En toute généralité, une transformation est une opération de modification d'un schéma (ou d'un algorithme) qui conserve certains aspects conceptuels, statistiques et techniques de ce schéma (ou d'un algorithme). Les transformations les plus intéressantes sont celles qui conservent tous les aspects. La notion de conservation est garantie par l'existence d'une transformation inverse qui permet de retrouver l'état initial du schéma transformé.

La gestion des quantifications (communément appelées statistiques) à travers les phases de conception pose trois problèmes : la représentation, la manipulation et l'exploitation des quantifications statiques et dynamiques. La représentation et la manipulation des statistiques doivent permettre la conception de base de données comme décrit plus haut, mais les transformations porteront en plus sur les statistiques. Ce qui fait que nous devrons parler de transformation de bases de données au lieu de transformation de schémas.

1.2. Etat de l'art

Tandis que la conception de bases de données fait depuis bien longtemps partie d'un domaine très pratiqué de l'informatique, la gestion informatisée des statistiques est, à ma connaissance, toujours au stade de recherche.

La gestion des statistiques est un des domaines de recherche de la Faculté d'Informatique ainsi que la validation de l'ensemble des contraintes structurelles d'une base de données. Deux références qui abordent le problème de la gestion des statistiques.

Le premier article *A temporal Statistical Model for Entity-Relationship Schemas* [HAI,92] présente les statistiques, les relations entre elles, les équations de transformations et une proposition de gestion des statistiques. La solution d'une gestion des statistiques fournit les valeurs des statistiques dérivables à partir d'un sous-ensemble de valeurs connus. Toutefois la gestion manque de souplesse parce que le choix du sous-ensemble des valeurs connus n'est pas totalement libre.

Le deuxième article *TRAMIS : a transformation-based database CASE tool* [HAI-CAD-DEC-MARb,92] présente l'atelier logiciel TRAMIS, et introduit brièvement les statistiques dans la conception de bases de données avec un exemple de transformation.

1.3. Objectifs et organisation du travail

La gestion des quantifications est limitée, dans le cadre du mémoire, à la représentation et la manipulation des quantifications statiques et vise surtout l'étude des problèmes qui surgissent pour une gestion souple des statistiques (qui laisse toute liberté au concepteur).

La première partie du travail introduit les statistiques dans la conception de base de données et est composée des chapitres 2 et 3.

Le chapitre 2. *Principes de la conception de bases de données* décrit un modèle de spécification, le modèle unique, ainsi que l'approche transformationnelle. L'atelier logiciel TRAMIS, un logiciel d'aide à la conception de bases de données, est brièvement présenté.

Le chapitre *3. Description statistique des données* décrit les statistiques d'une base de données correspondant aux concepts du modèle unique et l'ensemble des équations et des contraintes qu'elles respectent. Des transformations, les outils de travail de la conception de base de données, sont analysées dans leur aspect statistique.

La deuxième partie analyse les problèmes que pose la gestion des statistiques et analyse l'aspect intégration d'une solution logiciel TRAMIS : chapitres 4. et 5.

Le chapitre *4. Gestion des statistiques* analyse les problèmes et énonce des solutions que pose une gestion de statistiques. Une proposition de gestion est élaborée de manière abstraite.

Le chapitre *5. Application à l'environnement TRAMIS* présente deux propositions d'intégration, dans l'atelier logiciel TRAMIS, d'une gestion des statistiques d'une base de données. Les statistiques sont soit intégrées aux objets de la base de spécifications (structure actuelle de données) de TRAMIS, soit elles constituent un objet statistique. Les deux possibilités nécessitent des structures statistiques de gestion différentes, ainsi qu'une différente spécification dans la base de spécifications.

Chapitre 2

Principes de la conception de bases de données

Ce chapitre rappelle les concepts essentiels de la conception de bases de données. Pour la représentation du réel perçu (le premier objectif de la conception de base de donnée), le modèle unique utilisé permet de décrire l'aspect statique du réel perçu. Pour parvenir à la conception d'une gestion informatisée (le deuxième objectif de la conception de base de donnée), le concepteur utilise les transformations.

La démarche de conception peut être assistée par un outil logiciel. Nous allons découvrir avec TRAMIS, l'atelier logiciel de conception de bases de données conçu à l'Institut d'Informatique de Namur, l'aide qu'un atelier logiciel peut offrir.

2. Principes de la conception de bases de données

2.1. Le modèle unique

Le modèle d'expression d'une base de données que nous allons présenter est le modèle unique. C'est un modèle pratique parce qu'il couvre tout le processus de conception. Il est constitué de six objets de base : le schéma, l'entité, la relation, l'attribut, le groupe et l'espace.

Nous allons présenter sommairement les 6 objets du modèle unique. Les descriptions sont celles du manuel de référence de TRAMIS [CONCIS,90].

Schéma

Un schéma décrit un état déterminé de l'ensemble des structures d'une base de données. Il est constitué d'un nombre quelconque de type d'entités, type d'associations, d'attributs, de groupes et d'espaces.

Type d'entité

Un type d'entité correspond généralement à une classe d'objets de la réalité que les concepteurs reconnaissent comme ayant une importance majeure parmi les autres objets. On représentera graphiquement un type d'entité par un rectangle comportant en entête le nom du type d'entité.

Figure 2.1. : Les types d'entités OUVRAGE et AUTEUR

Les occurrences des types d'entités, les entités, sont les ouvrages et les auteurs.

Type d'association et rôles

Un type d'association correspond à une classe d'associations similaires entre des entités de la réalité, dont chacune joue un rôle déterminé. On représentera graphiquement un type d'association par un cartouche hexagonal comportant en entête le nom du type d'association et il est relié par des arcs éventuellement étiquetés du nom du rôle aux types d'entités membres.

Figure 2.2. : Le type d'association OUVAUT

Soit E un type d'entité jouant le rôle r1 dans le type d'association R. Le rôle r1 est caractérisé par une contrainte de cardinalité, constituée de deux entiers I-J tels que 0=I et I=J et qui spécifie que chaque occurrence de E joue de I à J fois le rôle r1 dans des occurrences de R. I et J sont dénommées respectivement cardinalité minimum et cardinalité maximum de r1. Lorsque I est égal à 0, le rôle est dit facultatif pour E. Dans le cas contraire, il est obligatoire. Lorsque J doit désigner un nombre arbitrairement grand (l'infini), on le représentera par la lettre N. On reportera les cardinalités sur la représentation graphique des relations.

Un type d'association possède au moins deux rôles. Le nombre de rôles s'appelle le degré du type d'association. Un type d'association de degré 2 est appelée binaire. Les relations binaires font l'objet d'une terminologie populaire. Soient r1, de cardinalité I1-J1 et r2, de cardinalité I2-J2, les deux rôles d'un type d'association.

Le type d'association est dite,

un-à-plusieurs si J1 = 1
plusieurs-à-un si J2 = 1
un-à-un si J1 = J2 = 1
plusieurs-à-plusieurs si J1 > 1 et J2 >1

Un type d'association fonctionnelle est un type d'association qui n'est pas plusieurs-à-plusieurs.

Attribut

Un attribut représente une propriété spécifique commune aux objets de la réalité. A chaque occurrence du type d'entité ou de type d'association à laquelle l'attribut est attaché, on associe un certain nombre (0,1, ou plusieurs) de valeurs de cet attribut. Dans la représentation graphique d'un schéma, le nom des attributs est indiqué dans le rectangle du type d'entité, ou dans la cartouche du type d'association.

A chaque attribut est associée une contrainte de cardinalité, constituée de deux entiers I-J tels que 0 = I et I = J et qui spécifie que chaque valeurs de l'attribut qui sont associées à chaque occurrence du type d'entité ou du type d'association doit être compris entre I et J.

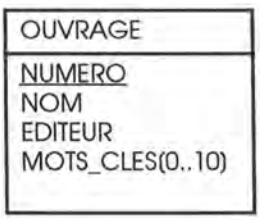


***Figure 2.3.** : L'attribut multi-valué MOTS_CLES*

Lorsque

I = 0, l'attribut est dit facultatif
I > 0, l'attribut est dit obligatoire
J = 1, l'attribut est dit mono-valué (ou répétitif)
I > 1, l'attribut est dit multi-valué

Un attribut décomposable est constitué d'autres attributs plus élémentaires. Les attributs non décomposables sont les attributs élémentaires. Dans la représentation graphique d'un schéma, la décomposition d'un attribut décomposable est indiquée par un décalage vers la droite du nom des composants.

***Figure 2.4. :** L'attribut décomposable LOCALISATION*

Un attribut élémentaire a un domaine de valeurs. Ce domaine est un ensemble de valeurs a priori. Ces valeurs sont caractérisées par leur type (ou format) et éventuellement par leur longueur.

Groupe

Un groupe est une collection d'attributs et /ou de rôles (ses composants), attachée à une entité ou une relation. Ce groupe joue un ou plusieurs rôles spécifiques pour cette entité ou cette relation : identifiant, clé d'accès ou référence. Nous ne proposons pas de représentation graphique unique pour la notion de groupe.

Groupe identifiant d'une entité

Si une collection d'attributs d'un type d'entité forme un groupe identifiant (ou plus communément, un identifiant) alors, étant donné une valeur de chacun de ces attributs, il ne peut à aucun instant exister plus d'une entité qui possède ces valeurs d'attributs.

La représentation graphique pour un groupe identifiant est facile à représenter s'il est composé uniquement d'attributs : nous les soulignons, sinon nous utilisons un cartouche aux bords arrondis.

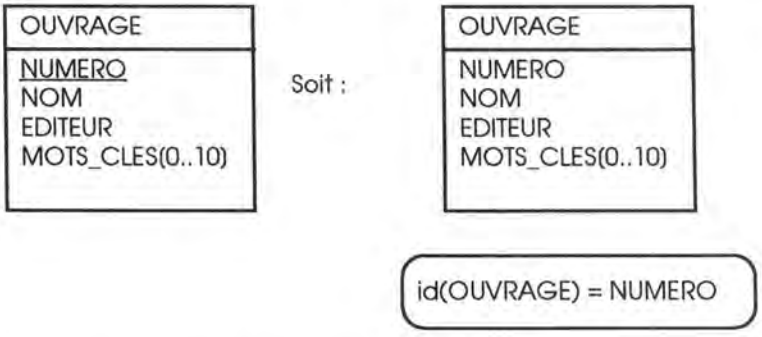


***Figure 2.5. :** Le groupe identifiant NUMERO*

Parmi les groupes identifiants d'un type d'entité, on en choisira arbitrairement un que l'on déclarera primaire. Les autres seront de ce fait secondaires. Cette notion répond au problème suivant : quel identifiant choisira-t-on de préférence lorsqu'il faudra identifier une occurrence d'un type d'entité ?

Groupe identifiant d'une relation

Un groupe d'identifiant d'une relation est constitué d'attributs et /ou de rôles de cette relation. Etant donné une valeur de chacun de ces attributs et d'une occurrence des types d'entités jouant chaque rôle, il ne peut, à tout instant, exister plus d'une relation ayant ces valeurs et associants ces entités.

Groupe clé d'accès d'une entité

Un groupe clé (d'accès) d'un type d'entité est constitué d'un ou plusieurs attributs et éventuellement d'un ou plusieurs rôles. Si un groupe clé est constitué d'attributs, alors étant donné une valeur de chaque attribut, il correspond un mécanisme d'accès qui permet d'accéder rapidement et sélectivement aux entités qui possèdent ces valeurs d'attribut.

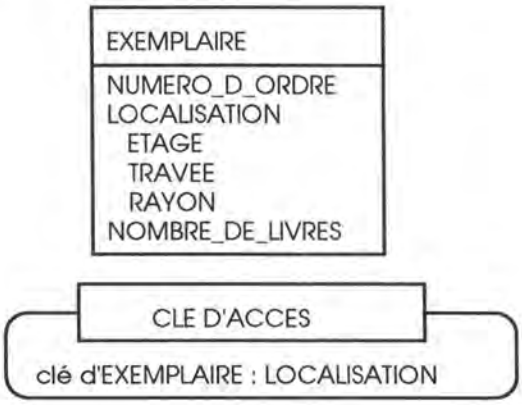


***Figure 2.6.** : Le groupe clé d'accès LOCALISATION*

Groupe de référence d'une entité

Un groupe de référence est une collection d'attributs (ou plus communément, attributs de référence) associées à un type d'entité, en relation avec un autre groupe de référence d'un type d'entité, différent ou non, tel que leurs composants soient comparables deux à deux. Cette structure définit une contrainte d'intégrité référentielle, laquelle peut prendre deux formes :

La première forme spécifie que les valeurs du groupe G du type d'entité E est une référence au type d'entité E', identifié par les valeurs du groupe G' ; ou encore que l'ensemble des valeurs de G de E **est inclus dans** l'ensemble des valeurs de G' de E'.

La deuxième forme spécifie que cette contrainte est valable dans l'autre sens: l'ensemble des valeurs de G' de E' est inclus dans l'ensemble des valeurs de G de E; par conséquent on dira aussi que l'ensemble des valeurs de G de E **est égal à** l'ensemble des valeurs de G' de E'.

Espace

Un espace est un regroupement d'un nombre quelconque d'entités. En principe, cette notion est une abstraction d'un espace de stockage dans lequel il est possible de ranger des occurrences des entités concernées.

Domaine

Un domaine permet de définir et de nommer des contraintes de valeurs (format et groupe de valeurs) associées à des attributs de types d'entités et de types d'associations. Les domaines permettent d'élargir indéfiniment la listes des types pré définis (numérique, alphanumérique, date, booléen).

2.2. Approche transformationnelle

L'approche transformationnelle de la conception de Système d'Informations et de Bases de Données, est l'élément essentiel de toute démarche. Nous allons revoir les principes de cette approche, agrémenté d'exemples de la référence [CONCIS, 90].

2.2.1. Transformation d'un type d'association en type d'entité

Principes

Un type d'association R est remplacé par un type d'entité de nom E et par un type d'association binaire R_i pour chaque rôle r_i du type d'association d'origine. Les cardinalités des rôles joués par le type d'entité E sont égales à 1-1, tandis que les cardinalités des autres rôles sont héritées des rôles r_i d'origine.

Exemple

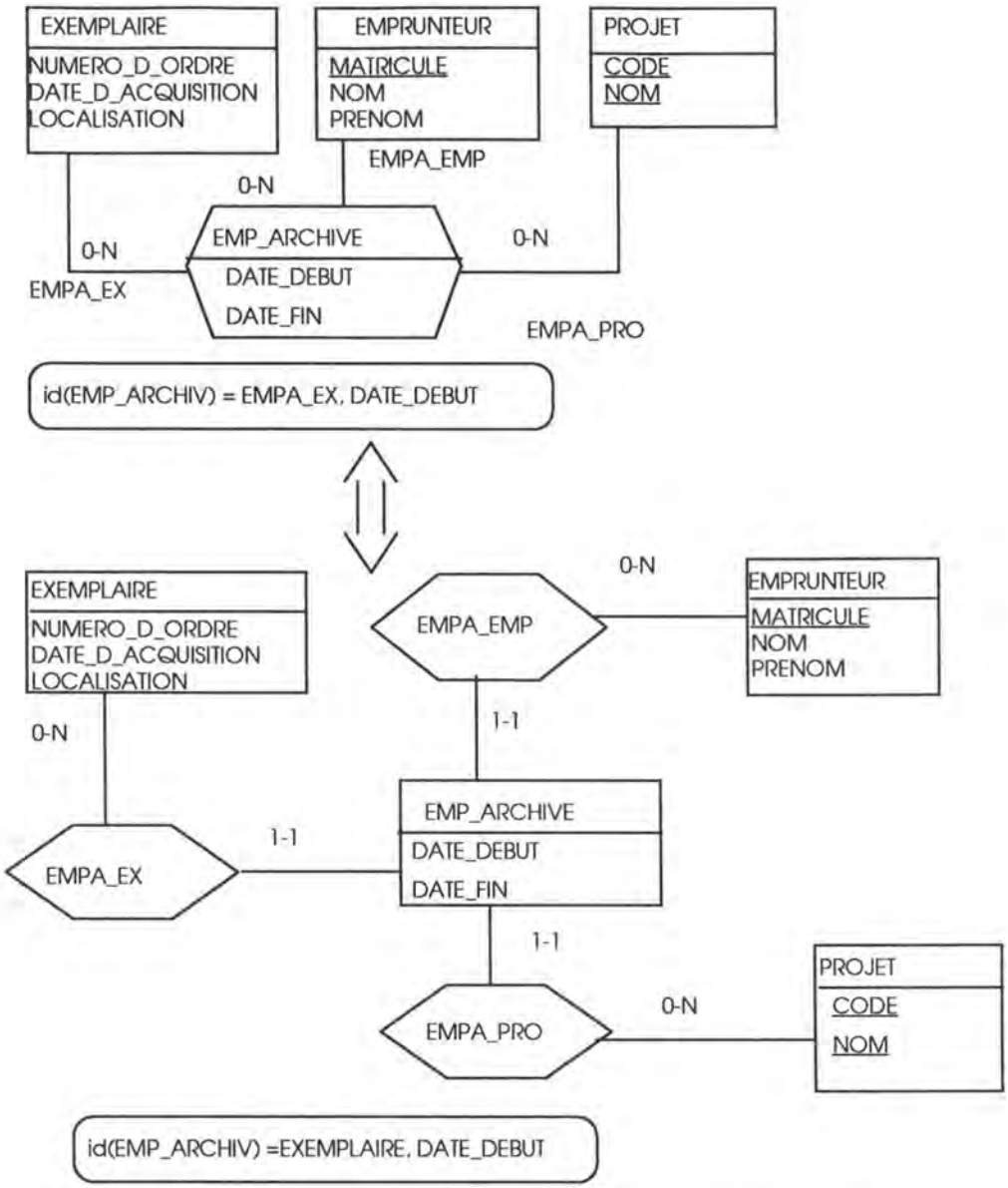


***Figure 2.7.** : Le type d'association EMP_ARCHIVE, qui représente les emprunts clôturés, réalisés chacun par un emprunteur pour le compte d'un projet, est remplacé par le type d'entité de même nom, de manière telle qu'à chaque occurrence de ce type d'association corresponde une occurrence du nouveau type d'entité. Chaque occurrence du type d'entité EMP_ARCHIVE est attachée à l'occurrence d'EXEMPLAIRE correspondant à l'exemplaire emprunté, à l'occurrence d'EMPRUNTEUR relative à l'emprunteur, et à l'occurrence de PROJET pour lequel l'emprunt a eu lieu. On observe la traduction de l'identifiant.*

Applications

- Eliminer une relation dont le degré est supérieur à 2.

On observe que la transformation ne produit que des types de relations binaires fonctionnelles, conforme à CODASYL par exemple.

- Eliminer une relation dotée d'attributs.

Les attributs de la relation sont alors attachés au type d'entité qui le remplace, conforme à CODASYL par exemple.

- Eliminer une relation binaire plusieurs-à-plusieurs.

Ces types de relations sont refusés dans de nombreux modèles, comme pour CODASYL par exemple.

- Eliminer une relation binaire récursive.

Les types de relation récursives sont refusés dans de nombreux modèles, comme pour CODASYL par exemple.

- Promotion d'une relation (au niveau conceptuel).

2.2.2. Transformation d'un attribut en type d'entité.

Principes

Un attribut A d'un type d'entité E est déplacé vers un nouveau type d'entité E' attachée à E par un type d'association dont les cardinalités sont fonction de la répétivité et du caractère obligatoire de l'attribut A. Lorsque l'attribut A est *répétitif* et *non identifiant*, on propose deux transformations distinctes. Selon la première, chaque occurrence de E' représente une valeur distincte de l'attribut A, **transformation par représentation de valeurs**. L'attribut déplacé est donc l'identifiant de E'. Selon la seconde, chaque occurrence de E' représente la présence dans une occurrence de E d'une valeur de A, **transformation par représentation d'instances**. Dans ce dernier cas, l'attribut déplacé n'est pas un identifiant de E'.

Exemple 1

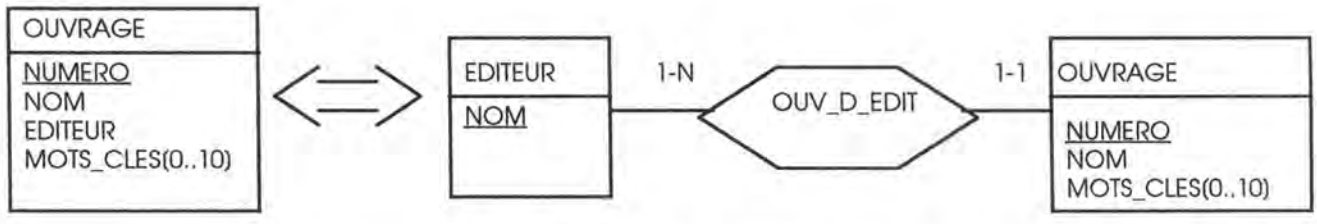


***Figure 2.8.** : L'attribut EDITEUR d'OUVRAGE est déplacé vers le nouveau type d'entité EDITEUR et renommé NOM. Chaque occurrence d'EDITEUR représente une valeur distincte de l'attribut d'origine. L'EDITEUR représente donc la liste des éditeurs ayant édité au moins un ouvrage répertorié. A chaque occurrence d'EDITEUR correspond une ou plusieurs occurrences d'OUVRAGE. Le concepteur peut généraliser le schéma en acceptant de représenter des éditeurs sans ouvrages (cardinalité 0-N au lieu de 1-N).*

Exemple 2

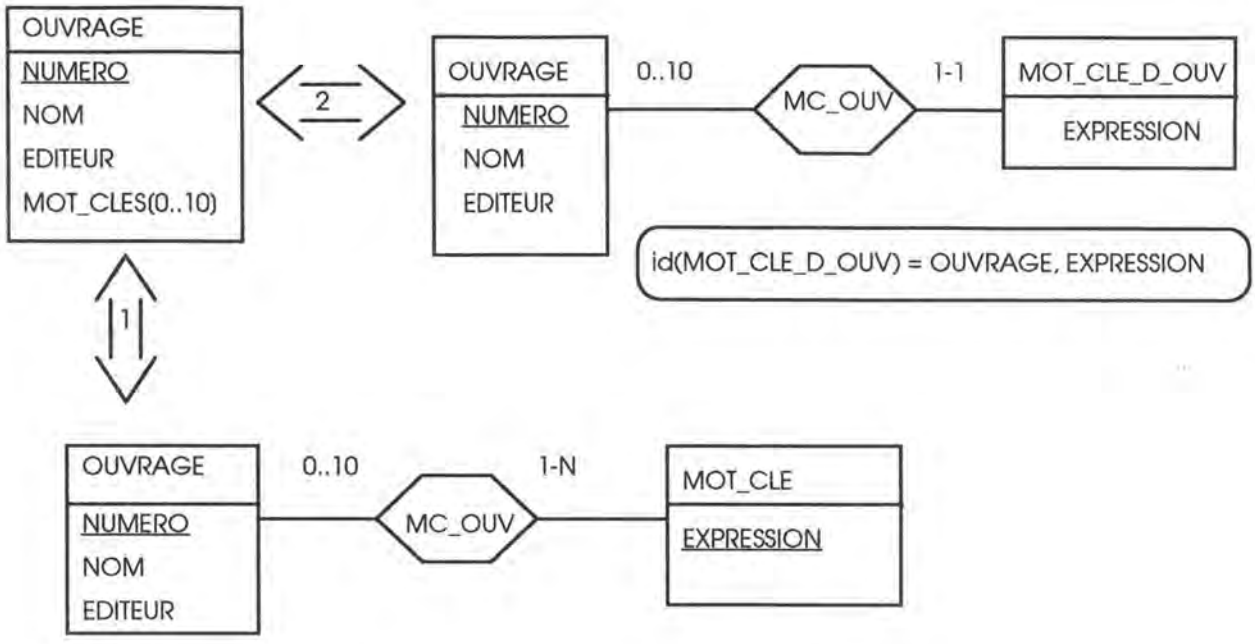


***Figure 2.9.** : On transforme ici l'attribut MOTS_CLES d'OUVRAGE (schéma supérieur), qui est répétitif et non identifiant. Dans ce cas, deux transformations sont possibles. (1) Dans la transformation du bas, on désire représenter la notion de mot-clé d'une manière autonome sous la forme du type d'entité MOT_CLE. Le type d'association MC_OUV est plusieurs-à-plusieurs et EXPRESSION (qui est l'attribut MOTS_CLES transféré et renommé) est l'identifiant du type d'entité MOT_CLE. (2) Dans la transformation du haut, on représente par une occurrence de MOT_CLE_D_OUV toute apparition d'une valeur de l'attribut MOT_CLES d'OUVRAGE. Le type d'association MC_OUV est à présent un-à-plusieurs, tandis que l'attribut EXPRESSION n'est pas un identifiant. Il constitue par contre avec le rôle OUVRAGE un identifiant de MOT_CLE_D_OUV, indiquant par là que les occurrences de MOT_CLE_D_OUV d'une même occurrence d'OUVRAGE ont des valeurs distinctes d'EXPRESSION.*

Applications

- Promotion d'un attribut en type d'entité.

Lors de l'élaboration d'un schéma conceptuel, il est fréquent qu'un simple attribut soit élevé au statut de type d'entité suite à la prise en compte de nouvelles spécifications : il apparaît par exemple que le concept représenté par l'attribut doit avoir lui aussi des attributs ou encore qu'il doit être associé à d'autres type d'entités.

- Eliminer un attribut répétitif.

Certains SGBD (relationnel par exemple) proposent un modèle qui n'accepte pas la notion d'attribut répétitif. En outre, un attribut répétitif peut-être une représentation maladroite d'un type d'association *un-à-plusieurs* ou *plusieurs-à-plusieurs* avec un type d'entité caché qu'il conviendrait de révéler. Cette transformation permet de *normaliser* un schéma conceptuel.

- Eliminer un attribut décomposable.

Selon le modèle de certains SGBD (relationnel par exemple) la notion d'attribut décomposable est ignorée. D'autre part, un attribut répétitif peut-être parfois considéré comme une représentation maladroite d'une entité associée. Cette transformation permet d'isoler l'attribut litigieux dans un type d'entité autonome.

- Eliminer un identifiant répétitif.

Malgré son intérêt conceptuel, la notion d'*attribut identifiant répétitif* est quasi partout exclue des modèles des SGBD, même si ceux-ci acceptent les attributs répétitifs (COBOL, CODASYL).

- Eliminer une clé d'accès répétitif.

La même restriction est plus souvent de mise en ce qui concerne clés d'accès répétitives.

- Eliminer une clé d'accès secondaire.

Certains SGBD (CODASYL) n'admettent, par type d'entité, qu'une seule clé d'accès constituée d'attributs. Cette transformation remplace un accès par clé par un accès via un chemin.

2.2.3. Transformation d'une relation en attributs de référence.

Principes

Cette transformation s'applique aux types de relations binaires fonctionnelles (dont un des rôles est de cardinalité 1-1 ou 0-1). Soit un type d'association, entre E1 et E2, telle que E1 y joue un rôle de cardinalité 1-1 ou 0-1, et telle que E2 possède un identifiant primaire qui n'est constitué que d'attributs. Ce type d'association est remplacé par des attributs de référence Ai, attachés à E1, qui sont des copies des attributs de l'identifiant de E2. Les nouveaux attributs font l'objet d'une contrainte d'intégrité référentielle.

Exemple

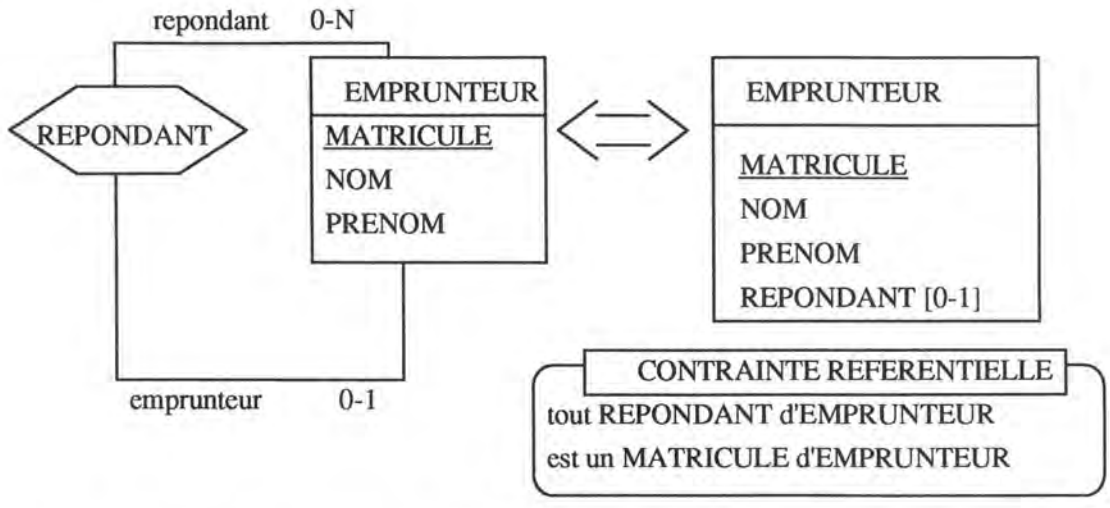


***Figure 2.10.** : Le type d'association REPONDANT est remplacé par l'attribut REPONDANT attaché à EMPRUNTEUR. Cet attribut est facultatif de fait de la cardinalité 0-1 de emprunteur.*

Applications

- Eliminer une relation dans un schéma COBOL ou relationnel.

Les modèles relationnels et les modèles de fichiers traditionnels du type COBOL ne disposent pas de structures de représentation directe de la notion de relation. Par combinaison des deux transformations, *relation → entité* et *relation → fonctionnelle attributs*, il est possible de transformer toute relation, quelle qu'en soit la complexité, en attributs de référence.

- Alléger un schéma conceptuel.

Lorsqu'en dernière analyse, on observe que le poids sémantique d'une entité est faible (par exemple, elle possède un seul attribut et participe à une seule relation), on peut envisager de diminuer son importance en la remplaçant par son identifiant dans les entités auxquelles elle est rattachée.

- Alléger un schéma CODASYL.

Dans un schéma CODASYL, remplacer une relation (un set type) peu utilisé par des attributs de référence peut dans certain cas diminuer le volume occupé, et le coût de gestion de cette relation.

2.2.4. Agrégation des attributs d'un groupe.

Principes

Une collection d'attributs (au moins un) est regroupée sous la forme d'un nouvel attribut décomposable. Les attributs candidats auront au préalable été rassemblés dans un groupe, qui après transformation ne contient plus que le nouvel attribut.

Exemple

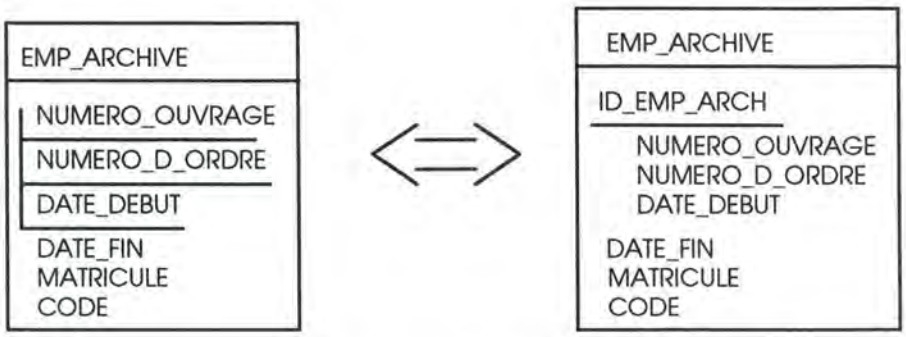


***Figure 2.11.** : Les trois composants de l'identifiant d'EMP_ARCHIVE sont rassemblés sous la forme de l'attribut décomposable ID_EMP_ARCH.*

Applications

- Restructurer les attributs d'un type d'entité ou d'un type d'association

Regrouper des attributs partageant des aspects sémantiques communs, partitionner les attributs en agrégats afin de réduire la complexité d'une longue liste d'attributs.

- Réduire un identifiant à un seul attribut

La présentation d'un identifiant multi-attribut sous la forme d'un seul composant peut simplifier la structure d'une entité.

- Réduire une clé d'accès à un seul composant

Dans certaines organisations de données on n'admet que des identifiants ou clés d'accès constitués d'un seul attribut (fichier COBOL séquentiels indexés par exemple).

2.2.5. Désagrégation d'un attribut décomposable.

Principes

Un attribut décomposable non répétitif est remplacé par ses composants.

Exemple

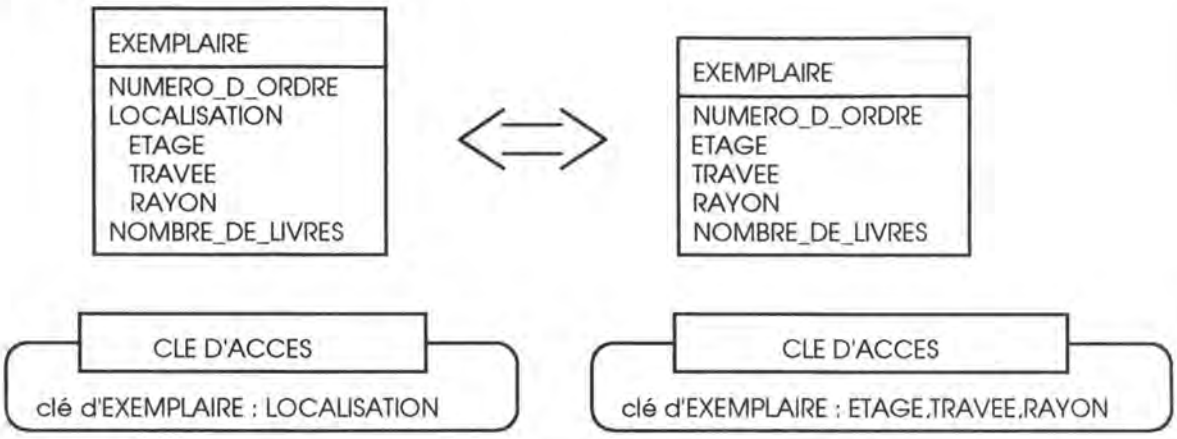


***Figure 2.12.** : L'attribut décomposable LOCALISATION est remplacé par ses composants ETAGE, TRAVEE, RAYON. Les concepts dans lesquels LOCALISATION intervient (ici une clé d'accès) sont ajustés en conséquence.*

Applications

- Promotion des constituants d'un attribut décomposable.

La désagrégation rend leur autonomie à des attributs qui avaient jusqu'ici été regroupés.

- Restructurer un attribut décomposable.

Une désagrégation, suivie d'une agrégation permet de redistribuer les composants d'attributs décomposables.

- Eliminer un attribut décomposable.

Certains modèles (relationnels par exemple) ne reconnaissent pas le concept d'attribut décomposable.

2.3. Cadre de travail TRAMIS

Le logiciel TRAMIS réalisé par l'Institut d'Informatique est destiné à aider les concepteurs de systèmes d'informations à analyser, concevoir et produire une base de données opérationnelle, correcte et efficace [le lecteur intéressé par une description détaillée se référera à la référence [RASE, 92]).

La description de la base de données en projet prend la forme d'un schéma, lequel est documenté, manipulé et analysé pour les besoins de la conception. Les schémas sont décrit avec un modèle de spécification unique, lequel couvre toutes les phases de conception. Ce modèle est constitué de six objets de base : le schéma, le type d'entité, le type d'association, l'attribut, le groupe et l'espace. (Ils sont présentés dans le chapitre suivant.) Le modèle permet différentes descriptions des objets d'une base de données : leur structure conceptuelle, leur sémantique, leurs propriétés statistiques ou encore leurs caractéristiques physiques. Le lecteur intéressé se référera à l'article [HAI-CAD-DEC-MAR,92] ou à l'article [HAI-CAD-DEC-MARb,92] qui les commentent brièvement.

Les schémas sont stockés dans la base de spécifications de TRAMIS. Nous pouvons appeler cela, la description statistique d'une base de données. Le lecteur intéressé se référera à [HAI-CAD-DEC-MAR,92] où l'on trouve une analyse détaillée.

Les fonctions de TRAMIS se présentent comme des outils indépendants opérants sur la base de spécifications. Les fonctions générales de l'atelier sont la documentation et la validation logique (phase de conception logique, conformité à un SGBD) et physique (phase de conception physique, conformité à un SGBD et exécutable).

La documentation regroupe la saisie, la consultation, la modification de spécifications et la production de la documentation d'analyses. La validation logique revient à l'utilisation pertinente des outils des transformations, tandis que la validation physique revient à la génération de code exécutable.

Chapitre 3

Description statistique des données

Ce chapitre présente la description statistique des données. La représentation du réel perçu permet de dégager des classes de données (type d'entités, type d'associations, ...). La description statistique correspond à des informations quantitatives (ou statistiques) de ces classes de données, que le concepteur possède. Les statistiques sont en relation les unes avec les autres.

Les transformations sont ensuite analysées pour leur aspect statistique.

3. Description statistique des données

Les statistiques d'une base de données constituent un ensemble de variables en relation les unes avec les autres. Dans ce chapitre, les statistiques sont présentées ainsi que leurs relations. Nous allons pour cela étendre le modèle unique avec la description statistique. L'article [HAI,92] présente ces statistiques et leur relations.

3.1. Description statique d'une base de données

3.1.1. Présentation des statistiques

La description statistique des objets d'une base de données renseigne sur les tailles des populations des entités, des relations et des groupes, ainsi que sur les longueurs et les populations des attributs.

Les statistiques sont regroupées selon les concepts du modèle Entité/Association qu'elles décrivent. Ces statistiques représentent de l'information sur l'extension (les occurrences) d'une base de données.

Type d'entité

- Population d'un type d'entité

  La taille (notée N_E) de la population d'un type d'entité (notée E) est le nombre moyen d'occurrences de ce type d'entité présentes dans la base de données à un instant de référence.

Type d'association

- Population d'un type d'association

  La taille (notée N_R) de la population d'un type d'association (notée R) est le nombre moyen d'occurrences de ce type d'association présentes dans la base de données à un instant de référence.

La notation $\mathbf{N_{ER}}$ spécifie la population d'un type d'entité ou d'un type d'association

Attribut

- Population d'un attribut

  La taille (notée N_A) de la population d'un attribut (noté A) est le nombre moyen de valeurs distinctes de cet attribut présentes dans la base de données à un instant de référence.

- La fréquence d'un attribut

  La fréquence moyenne (notée μ_A) d'un attribut (noté A) est le nombre moyen de valeurs associées aux occurrences d'un type d'entité (notée $\mu_{A/E}$) ou d'un type d'association (notée $\mu_{A/R}$).

  La notation $\boldsymbol{\mu_{A/ER}}$ spécifie la fréquence moyenne d'un type d'entité ou d'un type d'association.

- La probabilité d'une fréquence nulle d'un attribut

  La probabilité d'une fréquence nulle (notée Πo) d'un attribut (noté A) est la probabilité que le nombre moyen de valeurs associées aux occurrences d'un type d'entité (noté $\Pi o_{A/E}$)ou d'un type d'association (noté $\Pi o_{A/R}$) soit nulle.

  La notation $\mathbf{\Pi o_{A/ER}}$ spécifie la fréquence moyenne d'un type d'entité ou d'un type d'association.

- La fréquence non nulle d'un attribut

  La fréquence moyenne non nulle (notée μ'_A) d'un attribut (noté A) est le nombre moyen de valeurs associées aux occurrences d'un type d'entité (notée $\mu'_{A/E}$) ou d'un type d'association (notée $\mu'_{A/R}$) ayant au moins une valeur.

  La notation $\mathbf{\mu'_{A/ER}}$ spécifie la fréquence moyenne d'un type d'entité ou d'un type d'association.

- Longueur d'un attribut

  La longueur moyenne (notée λ_A) d'un attribut (noté A) est la longueur moyenne des valeurs de l'attribut A.

- Longueur non nulle d'un attribut

  La longueur moyenne non nulle (notée λ'_A) d'un attribut (noté A) est la longueur moyenne des valeurs de l'attribut A ayant au moins une valeur.

- L'attribut clé de fréquence

  L'attribut clé(notée A) comme fréquence moyenne (notée μ) d'une clé est le nombre moyen d'occurrences d'un type d'entité (notée $\mu_{E/A}$) ou d'un type d'association (notée $\mu_{R/A}$) pour une valeur de A.

  La notation $\mathbf{\mu_{ER/A}}$ spécifie l'attribut clé de fréquence d'un type d'entité ou d'un type d'association.

Rôle

- Cardinalité moyenne d'un rôle

  La cardinalité moyenne (notée μ_{ri}) d'un rôle (noté ri) d'un type d'association (notée R) est le nombre moyen d'occurrences de R dans lesquelles les occurrences du type d'entité jouent le rôle ri.

- La probabilité d'une cardinalité nulle d'un rôle

  La probabilité d'une cardinalité nulle (notée Πo) d'un rôle (noté ri) est la probabilité que le nombre moyen de valeurs associées aux occurrences d'un type d'entité (noté Πo_{ri}) soit nulle.

- La cardinalité non nulle d'un rôle

  La cardinalité moyenne non nulle (notée μ'_{ri}) d'un rôle (noté ri) est le nombre moyen de rôle associé aux occurrences d'un type d'entité (notée μ'_{ri}) jouant au moins un rôle.

Groupe

- Population d'un groupe

  La taille (notée N_G) de la population d'un groupe (noté G) est le nombre moyen d'occurrences de ce groupe associé à un type d'entité ou à un type d'association et correspond au nombre moyen de valeurs distinctes de ce groupe présentes dans la base de données à un instant de référence.

- Le groupe clé de fréquence

  Le groupe clé(notée G) comme fréquence moyenne (notée μ) d'une clé est le nombre moyen d'occurrences d'un type d'entité (notée $\mu_{E/G}$) pour une valeur de G.

- La fréquence d'un groupe

  La fréquence moyenne (notée μ) d'un groupe (notée G) est le nombre moyen d'occurrences d'un type d'entité (notée μ_G) pour une valeur de G.

Domaine

- Population d'un domaine

  La taille (notée N_D) de la population d'un domaine (notée D) est le nombre d'occurrences de ce domaine.

- Longueur d'une valeur d'un domaine

  La longueur (notée λ_D) d'une valeur du domaine (noté D) est la longueur des valeurs du domaine.

Espace

- L'occupation d'un espace

  L'occupation moyenne (notée μ) d'un espace (noté S) est le nombre moyen d'occurrences du type d'entité (noté E) associées à l'espace (notée $\mu_{E/S}$).

Les descriptions sont trop longues pour en parler facilement. Nous proposons des descriptions des statistiques abrégées dans la table ci-dessus.

Entité N_E : population moyenne d'un type d'entité E
Relation N_R : population moyenne d'un type d'association R
Attribut N_A : nombre de valeurs distinctes de l'attribut A $\mu_{A/E}$: nombre moyen de valeurs d'un attribut A pour chaque entité e $\Pi o_{A/E}$: probabilité qu'une entité n'ait pas de valeurs de l'attribut A $\mu'_{A/E}$: nombre moyen de valeurs de l'attribut A pour les entités e ayant des valeurs λ_A : longueur moyenne des valeurs distinctes d'un attribut A λ'_A : longueur moyenne des valeurs de l'attribut A pour les entités ayant des valeurs $\mu_{E/A}$: nombre moyen d'entités e correspondant à une valeur de l'attribut A
Rôle μ_{ri} : nombre moyen de fois qu'une entité e joue le rôle ri Πo_{ri} : probabilité qu'une entité e ne joue pas le rôle ri μ'_{ri} : nombre moyen de fois que le rôle ri est joué par les entités jouant le rôle
Groupe N_G : nombre moyen de valeurs distinctes de G μ_G : nombre moyen d'instances correspondant à une valeur de G
Domaine N_D : nombre de valeurs du domaine D λ_D : longueur moyenne des valeurs du domaine D
Espace $\mu_{e/S}$: nombre moyen d'entités e pour l'espace S

Table 3.1. : *Description des statistiques*

La description statistique est rajoutée dans un cartouche aux bords arrondis. Le petit exemple suivant (peu d'objets mais beaucoup de statistiques !) illustre les propos sur les descriptions :

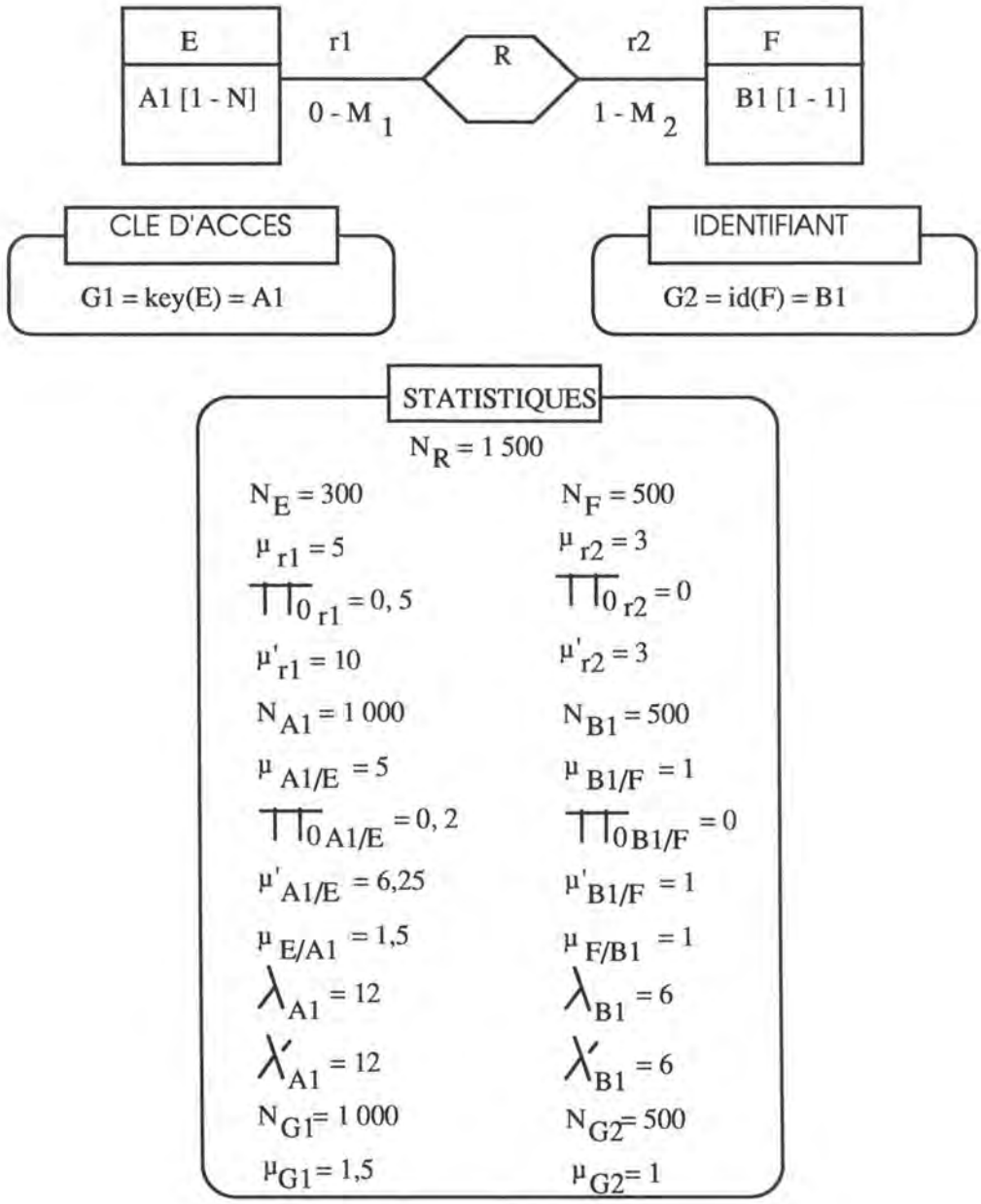


***Figure 3.1.** : Illustration des descriptions statistiques*

3.1.2. Modèle statique des statistiques

a) Les types d'équations

Nous allons agrémenter les types d'équation avec l'exemple suivant :

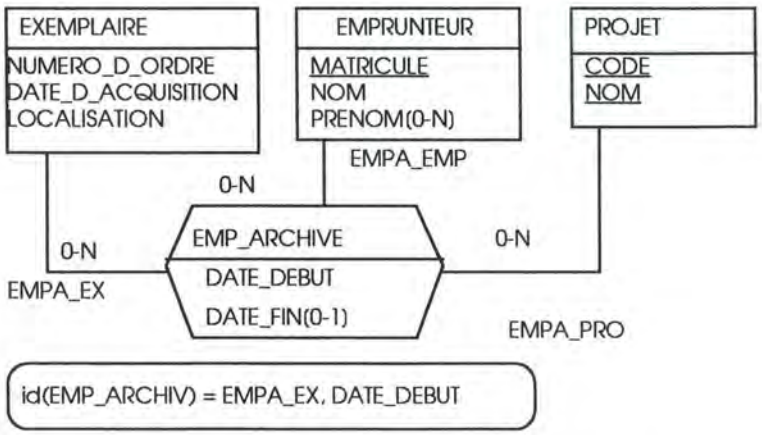


Figure 3.2. : *Illustration des types d'équations*

T1. $N_R = N_{Ei} \times \mu_{ri}$

Exemple : Considérons la figure 2.7.

$N_{EMP\_ARCHIVE} = N_{EXEMPLAIRE} \times \mu_{EMPA\_EX}$
$N_{EMP\_ARCHIVE} = N_{EMPRUNTEUR} \times \mu_{EMPA\_EMP}$
$N_{EMP\_ARCHIVE} = N_{PROJET} \times \mu_{EMPA\_PRO}$

T2. $N_{ER} \times \mu_{A/ER} = N_A \times \mu_{ER/A}$

Exemple : Considérons la figure 2.7.

$N_{PROJET} \times \mu_{CODE/PROJET} = N_{CODE} \times \mu_{PROJET/CODE}$
$N_{EMP\_ARCHIVE} \times \mu_{DATE\_FIN/EMP\_ARCHIVE} = N_{DATE\_FIN} \times \mu_{EMP\_ARCHIVE/DATE\_FIN}$

T3. $\mu'_{ri} = \mu_{ri} / (1 - \Pi o_{ri})$

Exemple : Considérons la figure 2.7.

$\mu'_{EMPA\_EX} = \mu_{EMPA\_EX} / (1 - \Pi o_{EMPA\_EX})$

T4. $\mu'_{A/ER} = \mu_{A/ER} / (1 - \Pi o_{A/ER})$

$\mu'_{PRENOM/EMPRUNTEUR} = \mu_{PRENOM/EMPRUNTEUR} / (1 - \Pi o_{PRENOM/EMPRUNTEUR})$

$\mu'_{DATE\_FIN/EMP\_ARCHIVE} = \mu_{DATE\_FIN/EMP\_ARCHIVE} / (1 - \Pi o_{DATE\_FIN/EMP\_ARCHIVE})$

T5. $\lambda'_A = \lambda_A / (1 - \Pi o_{A/ER})$

$\lambda'_{PRENOM/EMPRUNTEUR} = \lambda_{PRENOM/EMPRUNTEUR} / (1 - \Pi o_{PRENOM/EMPRUNTEUR})$

$\lambda'_{DATE\_FIN/EMP\_ARCHIVE} = \lambda_{DATE\_FIN/EMP\_ARCHIVE} / (1 - \Pi o_{DATE\_FIN/EMP\_ARCHIVE})$

b) Les constantes et les équations sous contrainte structurelle

- Rôle

$m_{ri} > 0 \Rightarrow \Pi o_{ri} = 0$

$m_{ri} = M_{ri} \Rightarrow \mu_{ri} = m_{ri}$

$m_{ri} = 0$ et $M_{ri} = 1 \Rightarrow \mu'_{ri} = 1$

(Cette relation est dérivée de la suivante : $\Pi o_{ri} = 1 - \mu_{ri}$)

relation récursive de rôles r1 et r2 $\Rightarrow$ T6. $\mu_{ri} = \mu_{rj}$

- Attribut

$m_A > 0 \Rightarrow \Pi o_{A/ER} = 0$

$m_A = M_A \Rightarrow \mu_{A/ER} = m_A$

$m_A = M_A = 1 \Rightarrow \mu_{A/ER} = 1$

décomposable $\Rightarrow$ T7. $\lambda_A = \Sigma_i \lambda_{Ai}$

- Domaine

longueur fixe L $\Rightarrow \lambda_D = \lambda_A = L$

- Groupe

G a un seul composant A et est identifiant

$\Rightarrow N_A = N_G = N_{ER}$

$\Rightarrow \mu_{ER/A} = \mu_G = \mu_{A/ER} = \mu'_{A/ER} = 1$

$\Rightarrow \Pi o_{A/ER} = 0$

G a un seul composant A et n'est pas identifiant

$\Rightarrow N_A = N_G$

$\Rightarrow \mu_{ER/A} = \mu_G$

G identifiant et plusieurs composants

$\Rightarrow N_{ER} = N_G$

$\Rightarrow \mu_G = 1$

Contrainte d'égalité entre groupes

$\Rightarrow N_{G1} = N_{G2}$

$\Rightarrow \mu_{G1} = \mu_{G2}$

c) Les contraintes structurelles

- Les probabilités

$0 <= \Pi o_{A/ER} <= 1$

$0 <= \Pi o_{ri} <= 1$

- Attribut

$N_A <= N_D$

$m_A <= \mu_{A/E} <= M_A$

$m_A <= \mu'_{A/E} <= M_A$

L'attribut décomposable A, composants Ai :

$\forall i : \lambda_{Ai} <= \lambda_A$

- Rôle

$m_{ri} <= \mu_{ri} <= M_{ri}$

$m_{ri} <= \mu'_{ri} <= M_{ri}$

- Groupe

$0 <= N_G <= N_E$

3.2. Statistiques d'évolution

Ce mémoire étend la gestion des statistiques à une évolution temporelle linéaire des statistiques. Appelons le schéma, avec ses valeurs statiques, le schéma initial à l'instant T_o". Si on considère un schéma courant à l'instant T_i, on peut exprimer l'évolution de la statistique s_i de ce schéma en fonction de sa valeur s_{io} à l'instant T_o et de l'incrément $d(s_i)$ par période T de la statistique.

$$s_i = s_{io} + d(s_i) \times (T_i - T_o) / T$$

Toutes les statistiques courantes d'un schéma sont caractérisées par l'instant T_i et par une même période T. L'incrément par période $d(s_i)$ est indépendant pour chaque statistique.

3.3. Transformations des statistiques

Des relations particulières entre statistiques de bases de données (transformées) sont analysées dans ce chapitre. Les statistiques des nouvelles structures sont dérivées à partir des composants de la structure d'origine.

3.3.1. Transformation d'un type d'association en type d'entité

Description

Un type d'association est transformé en un type d'entité et chacun de ses rôles en type d'association et inversement.

Définition

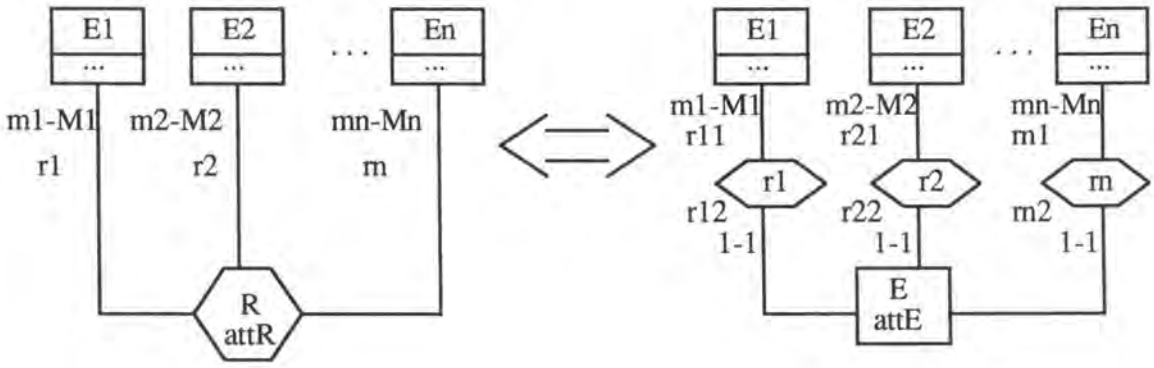


Figure 3.3. : Transformation d'un type d'association en type d'entité

Les statistiques conservées :

$$N_{E1}, N_{E2}, \ldots N_{En}$$

Les statistiques des attributs attR sont les mêmes que les statistiques des attributs attE

Les statistiques déduites de la structure :

$$\mu_{rk2} = 1 \qquad \forall i : 1 =< k =< n$$
$$\Pi o_{rk2} = 0 \qquad \forall i : 1 =< k =< n$$
$$\mu'_{rk2} = 1 \qquad \forall i : 1 =< k =< n$$

Les statistiques transformées :

$N_R = N_E$

(Les rôles sont renommés :)

$\mu_{ri} = \mu_{rk1}$ $\forall i : 1 =< i = k =< n$

$\Pi o_{ri} = \Pi o_{rk1}$ $\forall i : 1 =< i = k =< n$

$\mu'_{ri} = \mu'_{rk1}$ $\forall i : 1 =< i = k =< n$

$N_E = N_R$

$N_{ri} = N_R$ $\forall i : 1 =< i =< n$

(Les rôles sont renommés :)

$\mu_{rk1} = \mu_{ri}$ $\forall i : 1 =< i = k =< n$

$\Pi o_{rk1} = \Pi o_{ri}$ $\forall i : 1 =< i = k =< n$

$\mu'_{rk1} = \mu'_{ri}$ $\forall i : 1 =< i = k =< n$

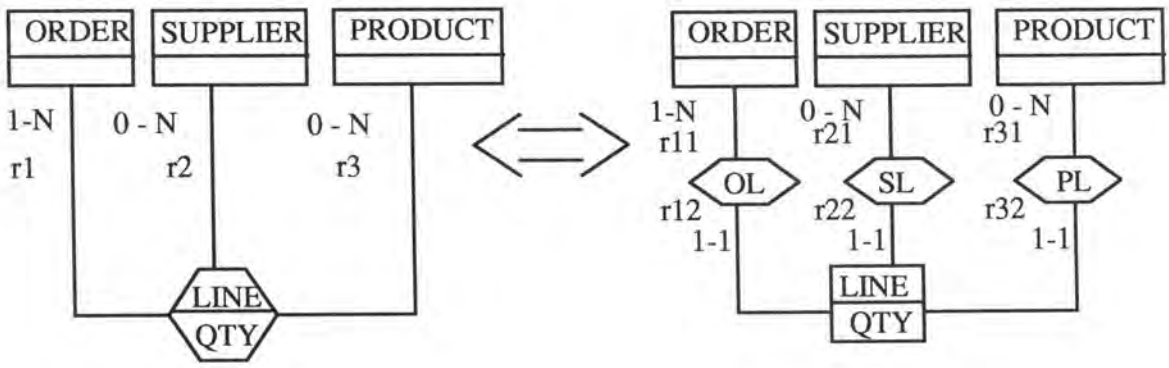
Tous les rôles font partie de l'identifiant G de E

$N_G = N_R$, $\mu_G = 1$

Le rôle r_i est l'identifiant G de E

$N_G = N_i \times (1 - \Pi o_{ri})$, $\mu_G = 1$

Nous trouvons un exemple d'une transformation de la description statistique de ce type dans l'article [HAI, 92].



$N_{ORDER} = 2\,000$

$N_{SUPPLIER} = 300$

$N_{PRODUCT} = 30\,000$

$\mu_{r1} = 3$

$\mu_{r2} = 50$

$\mu_{r3} = 0,2$

$N_{ORDER} = 2\,000$

$N_{SUPPLIER} = 300$

$N_{PRODUCT} = 30\,000$

$\mu_{r11} = 3$ $\mu_{r12} = 1$

$\mu_{r21} = 50$ $\mu_{r22} = 1$

$\mu_{r31} = 0,2$ $\mu_{r32} = 1$

$N_{LINE} = 6\,000$

Figure 3.3. : *Exemple de transformation d'un type d'association en type d'entité*

Vérifions les calculs :

Les statistiques conservées :

N_{ORDER} , $N_{SUPPLIER}$ et $N_{PRODUCT}$

Les statistiques déduites de la structure :

μ_{r11} , μ_{r21} et μ_{r31}

Les statistiques transformées :

μ_{r1} , μ_{r2} et μ_{r3} sont renommées μ_{r11} , μ_{r21} et μ_{r31}.

La dernière statistique, la N_{LINE} ne peut être obtenue directement avec l'équation $N_E = N_R$, nous n'avons pas la statistique N_{LINE}.

Nous allons utiliser l'équation que les statistiques doivent respecter $N_R = N_E \times \mu_{ri}$. (Rappelons que ces équations sont commentées dans la section 3.1.2. Modèle statique des équations.)

$N_{LINE} = N_{ORDER} \times \mu_{r11} = 2\,000 \times 3 = 6\,000$

3.3.2. Transformation d'un attribut en type d'entité

A. Transformation par représentation de valeurs

Description

Un attribut d'un type d'entité est transformé en type d'entité, dont chaque entité représente une valeur distincte de l'attribut et qui est relié à son type d'entité d'origine.

Définition

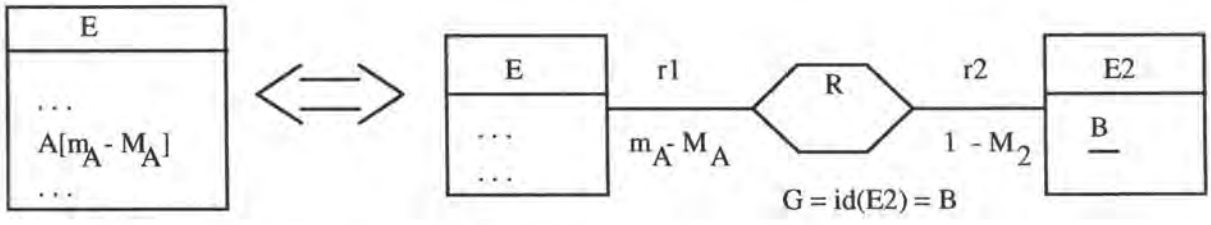


Figure 3.4. : Transformation par représentation de valeurs

Les statistiques conservées : N_E

Les statistiques déduites de la structure :

$\Pi o_{r2} = 0$
$\Pi o_{B/E} = 0$
$\mu_{B/E} = 1$
$\mu'_{B/E} = 1$
$\mu_{E/B} = 1$

$\mu_G = 1$

Les statistiques transformées :

$\mu_{A/E} = \mu_{r1}$
$\Pi o_{A/E} = \Pi o_{r1}$
$\mu'_{A/E} = \mu'_1$
$N_A = N_B$
$\lambda_A = \lambda_B$
$\mu_{E/A} = \mu_{r2}$
$\lambda'_A = \lambda_B / (1-\Pi o_{r1})$

$\mu_{r1} = \mu_{A/E}$
$\Pi o_{r1} = \Pi o_{A/E}$
$\mu'_{r1} = \mu'_{A/E}$
$\mu_{r2} = \mu_{E/A}$
$\mu'_{r2} = \mu_{E/A}$
$N_B = N_A$
$N_{E2} = N_A$
$N_G = N_A$
$\lambda_B = \lambda_A$
$\lambda'_B = \lambda_A$
$N_R = N_E \times \mu_{A/E}$

Nous trouvons un exemple d'une transformation de la description statistique de ce type dans l'article qui ne cite que brièvement la description statistique d'objets [HAI-CAD-DEC-MARb,92].

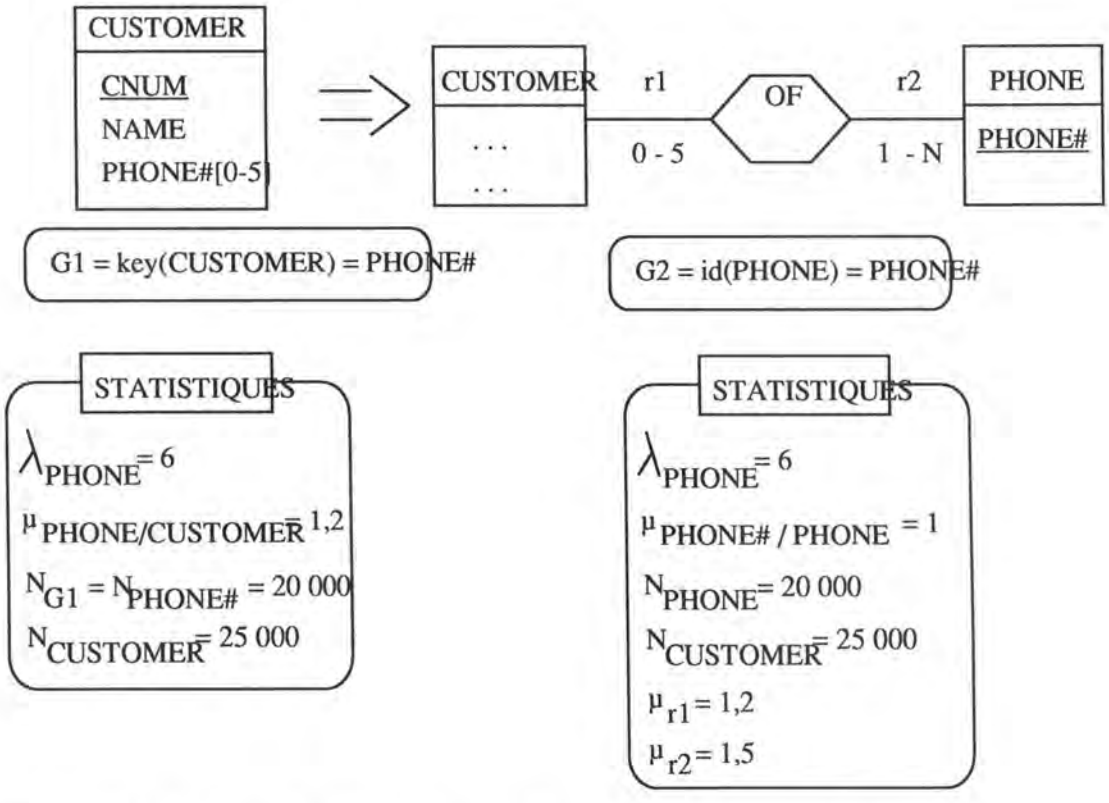


Figure 3.5. : *Exemple de transformation par représentation de valeurs*

Vérifions les calculs :

Les statistiques conservées :

$N_{CUSTOMER} = N_{CUSTOMER} = 25\,000$

Les statistiques transformées :

$\mu_{r1} = \mu_{PHONE/CUSTOMER} = 1,2$
$\mu_{r2} = N_{PHONE\#\ /\ CUSTOMER}$
$N_{PHONE} = N_{PHONE\#} = 20\,000$
$\lambda_{PHONE} = \lambda_{PHONE\#} = 6$
$\mu_{PHONE\#\ /\ PHONE} = 1$

La dernière statistique μ_{r2} ne peut être obtenue directement, nous n'avons pas la statistique $N_{PHONE\#\ /\ CUSTOMER}$. Nous allons utiliser l'équation $N_R = N_E \times \mu_{ri}$ que les statistiques doivent respecter.

$N_{OF} = N_{CUSTOMER} \times \mu_{PHONE\#/CUSTOMER}$
$= 25\,000 \times 1,2 = 30\,000$

$\mu_{r2} = N_{OF} / N_{PHONE} = 30\,000 / 20\,000 = 1,5$

B. Transformation par représentation d'instances

Description

Un attribut d'un type d'entité est transformé en type d'entité, dont chaque entité représente une instance de l'attribut et qui est relié à son type d'entité d'origine.

Définition

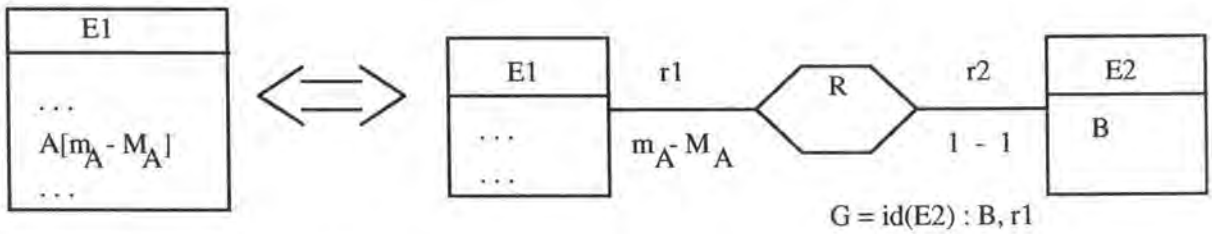


Figure 3.6. : *Transformation par représentation d'instances*

Les statistiques conservées :

N_{E1}

Les statistiques déduites de la structure :

$\mu_{r2} = 1$
$\Pi o_{r2} = 0$
$\mu'_{r2} = 1$

$\Pi o_{B/E2} = 0$
$\mu_{B/E2} = 1$
$\mu'_{B/E2} = 1$
$\mu_G = 1$

Les statistiques transformées :

$\mu_{A/E} = \mu_{r1}$
$\Pi o_{A/E} = \Pi o_{r1}$
$\mu'_{A/E} = \mu'_{r1}$

$N_A = N_B$
$\lambda_A = \lambda_B$
$\mu_{E/A} = \mu_{E2/B}$

$\lambda'_A = \lambda_B / (1 - \Pi o_{r1})$

$\mu_{r1} = \mu_{A/E}$
$\Pi o_{r1} = \Pi o_{A/E}$
$\mu'_{r1} = \mu'_{A/E}$

$N_B = N_A$
$\lambda_B = \lambda_A$, $\lambda'_B = \lambda_A$
$\mu_{E2/B} = \mu_{E/A}$

N_R , N_{E2} , $N_G = N_E \times \mu_{A/E}$

Nous trouvons un exemple d'une transformation de la description statistique de ce type dans l'article [HAI, 92].

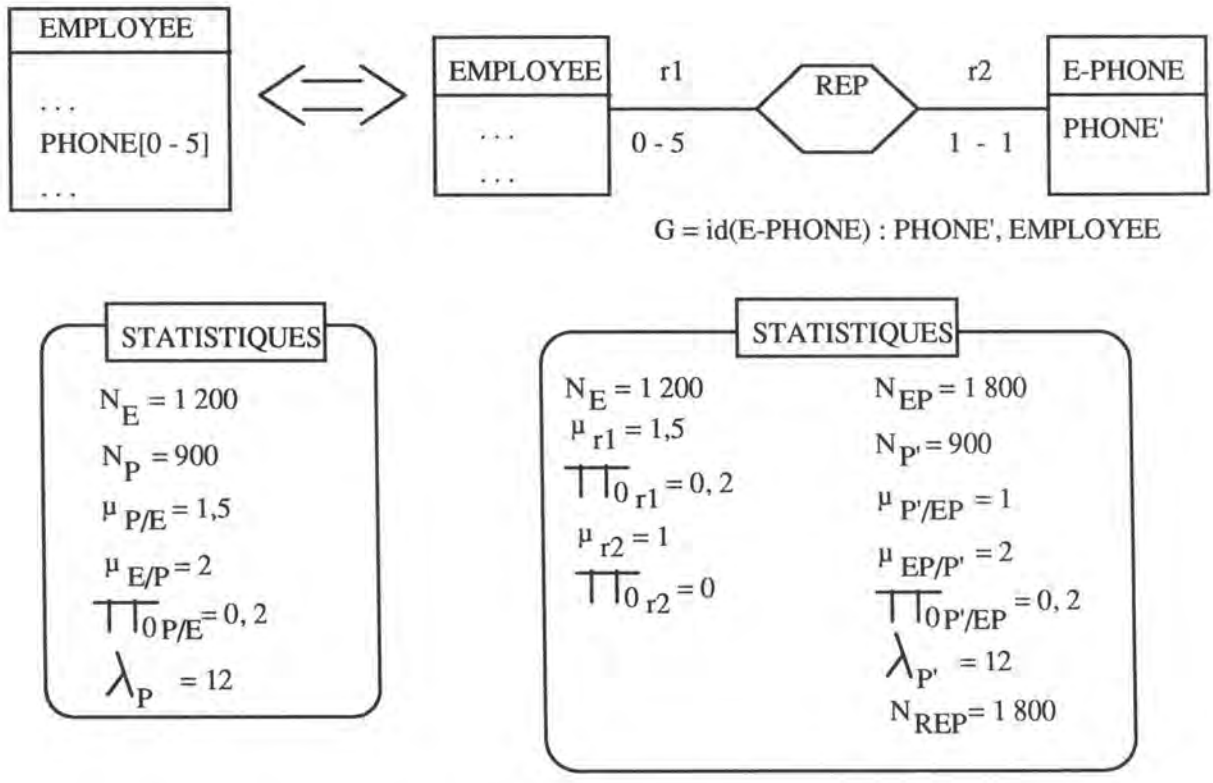


Figure 3.6. : *Exemple de transformation par représentation d'instances*

3.3.3. Transformation d'un type d'association en groupe de référence

Description

Un type d'association 1-N ou 1-1 entre deux types d'entité est remplacé par un groupe de référence, composé d'attributs.

Définition

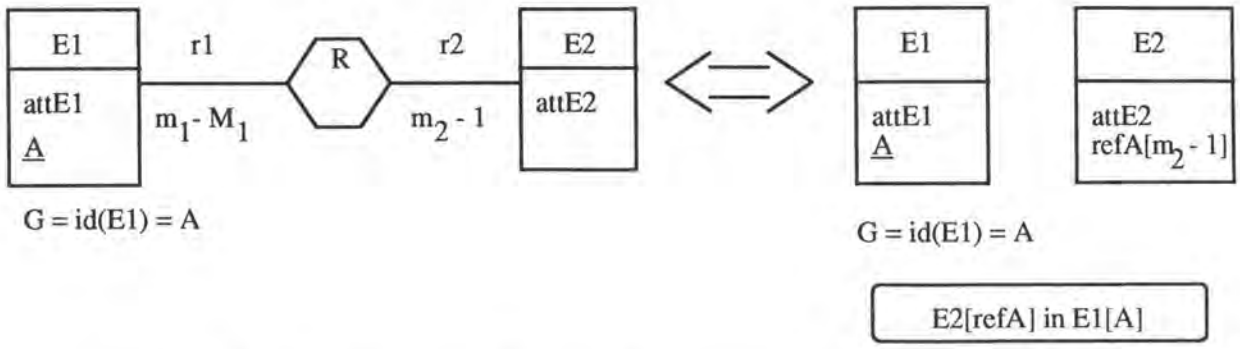


Figure 3.7. : Transformation d'un type d'association en groupe de référence

A est l'ensemble des attributs d'un identifiant du type d'entité E1

Les statistiques conservées :

N_{E1}, N_{E2}, les statistiques des attributs A, attE1 et attE2

Les statistiques déduites de la structure :

$\Pi o_{r1} = 0$

$\mu_{G1} = 1$

Les statistiques transformées :

$\mu_{r2} = \mu_{refA/E2}$
$\Pi o_{r2} = \Pi o_{refA/E2}$
$\mu'_{r2} = \mu'_{refA/E2}$
$\mu_{r1} = \mu_{E2/refA}$
$\Pi o_{r1} = 1 - N_{refA} / N_A$
$\mu'_{r1} = \mu_{E2/refA}$
$N_R = N_{E2} \times \mu_{refA/E2}$

$\mu_{refA/E2} = \mu_{r2}$
$\Pi o_{refA/E2} = \Pi o_{r2}$
$\mu'_{refA/E2} = \mu'_{r2}$
$N_{refA} = N_A \times (1-\Pi o_{r1})$
$\lambda_{refA} = \lambda_A \times (1-\Pi o_{r2})$
$\lambda'_{refA} = \lambda_A$
μ_{G2} , $\mu_{E2/refA} = \mu_{r1} / (1-\Pi o_{r1})$
N_{G2} , $N_{G1} = N_{E1}$

Remarque : Si $M_1 = 1$ alors au moins un des composants de refA est obligatoire, puisque refA devient identifiant du type d'entité E2.

Nous trouvons un exemple d'une transformation de la description statistique de ce type dans l'article [HAI, 92].

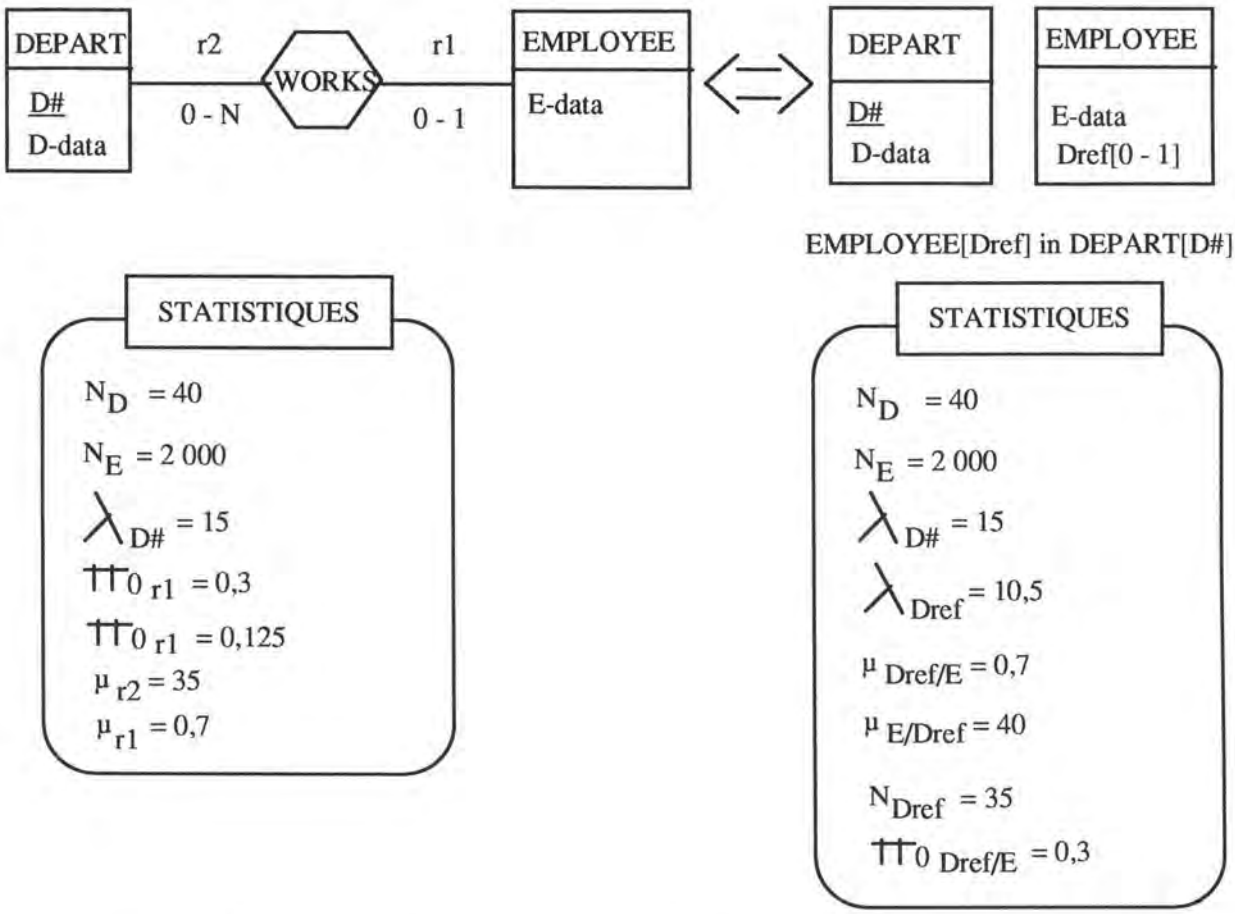


***Figure 3.8.** : Exemple de transformation d'un type d'association en groupe de référence*

Vérifions les calculs :

Les statistiques conservées :
N_D et N_E
Les statistiques des attributs D#, D-data, E-data

Les statistiques transformées :
$\mu_{Dref/E} = \mu_{r2} = 0,7$
$\Pi o_{Dref/E} = \Pi o_{r2} = 0,3$

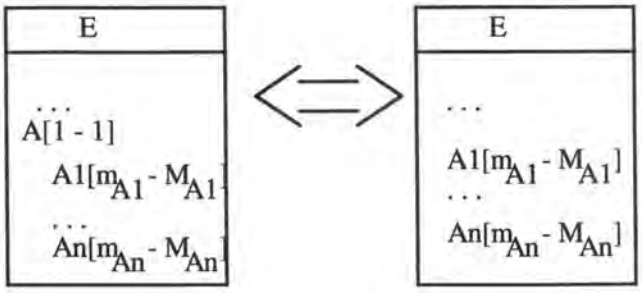
$N_{Dref} = N_{D\#} \times (1-\Pi o_{r1}) = 40 \times (1-0,125) = 35$
$\lambda_{Dref} = \lambda_{D\#} \times (1-\Pi o_{r2}) = 15 \times (1- 0,3) = 10,5$
$\mu_{G2}, \mu_{E/Dref} = \mu_{r1} /(1-\Pi o_{r1}) = 35/(1- 0,125) = 40$

3.3.4. Désagrégation d'un attribut décomposable

Description

Un attribut décomposable est remplacé par ses attributs composants.

Définition



Les statistiques conservées :

N_E

Les statistiques des attributs A1, ... An

Les statistiques déduites de la structure :

$\mu_{A/E} = 1$
$\Pi o_{A/E} = 0$
$\mu'_{A/E} = 1$

Les statistiques transformées :

$\lambda_A = \sum_{i=1}^{n} \lambda_{Ai}$
$\lambda'_A = \sum_{i=1}^{n} \lambda'_{Ai}$
$N_A = (N_E)^n / \prod_1^n \mu_{E/Ai}$
$\mu_{E/A} = \prod_1^n \mu_{E/Ai} / (N_E)^{n-1}$

Ces deux dernières équations sont à utiliser avec prudence, parce qu'elles exigent que les probabilités $\frac{\mu_{E/Ai}}{N_E}$ (les attributs Ai) soient indépendants.

Chapitre 4

Gestion des statistiques

Ce chapitre va au coeur du problème : il propose une gestion souple des statistiques, avec une démarche multi-niveaux, qui peut être perçue comme des états d'évolution successifs (processus) d'un modèle statistique : d'abord, la construction d'un **modèle statistique de classement** fixe le sens des équations, en fixant des statistiques comme constantes (appelées les statistiques de base); ensuite, des valeurs vérifiant les contraintes structurelles sont introduites pour les statistiques de base, et on aboutit dans le **modèle statistique des valeurs de base**; ensuite après dérivation on obtient le **modèle statistique des valeurs**; si ces valeurs vérifient les contraintes structurelles, le modèle est appelé le **modèle statistique des valeurs valides**, sinon des nouvelles valeurs de statistiques de base doivent être introduites et le processus est recommencé en partie.

Les valeurs du modèle statistique sont en plus dynamiques, tout comme les données de l'extension.

Les transformations doivent préserver les statistiques lorsque la base de donnée correspondante est transformée.

4. Gestion des statistiques

4.1. Introduction

Jusqu'à présent, le modèle statistique est un moyen de spécifier les statistiques d'une base de données, et les équations et contraintes qu'elles font intervenir. Appelons ce modèle : le **modèle statistique de spécification.**

La spécification statistique d'un schéma selon ce modèle donne un **modèle statistique des équations.** Ce modèle statistique des équations est un système d'équations (un ensemble complexe de contraintes), dont les variables sont les statistiques du schéma. La gestion des statistiques va faire évoluer le modèle statistique des équations vers une solution, qui correspond à une valeur pour chaque statistique, laquelle doit respecter l'ensemble des contraintes. Pour faciliter la compréhension nous allons donner des noms aux différents états d'évolution.

Le modèle statistique des équations forme un système d'équations dans lequel il y a plus de variables que d'équations. La construction d'un **modèle statistique de classement** consiste à fixer certaines de ces variables comme constantes et d'établir quelles variables sont dérivables. La gestion introduit la notion de *statistique de base* (statistique dont la valeur est supposée connue par l'utilisateur) pour les variables fixées comme constantes et de *statistique dérivée* (dérivable de celles connues) pour les variables dérivables. Ainsi les statistiques d'un schéma sont à classer en ces termes à la construction du modèle statistique de classement par lequel le concepteur précise ce qu'il apporte comme valeurs et par conséquent quelles statistiques sont dérivables.

Le **modèle statistique des valeurs de base** correspond au modèle statistique de classement où les statistiques de base ont reçu une valeur. Les valeurs des statistiques de base respectent les contraintes structurelles.

L'ensemble des valeurs des statistiques d'un schéma forme le **modèle statistique des valeurs**, lequel respecte le modèle statistique d'équations (où les valeurs vérifient les équations) et le modèle statistique des valeurs de base, où une statistique dérivée correspond à une valeur dérivée. Si toutes les contraintes structurelles sont respectées alors nous avons **modèle statistique des valeurs valides**.

- Remarque : Rappelons qu'à tout instant il n'y a qu'un modèle statistique. Les différents modèles statistiques que nous venons de décrire correspondent à des états d'évolution différents.

La distinction de modèle statistique de classement et de modèle statistique des valeurs offre beaucoup de souplesse : Il y a la possibilité du choix des statistiques de base au départ, mais aussi la possibilité de modifier le modèle statistique de classement sans pour cela perdre le modèle totalement (ni tout le classement, ni toutes les valeurs). Une façon plus rigide d'agir est d'obliger l'utilisateur d'introduire les valeurs de certaines statistiques et de calculer les valeurs des autres.

4.2. Un exemple de construction d'un modèle statistique de classement

L'exemple qui suit permet d'aborder plus facilement la construction d'un modèle statistique de classement et de localiser les problèmes que l'on rencontre. L'exemple soulève des questions et éclaire des propriétés d'un modèle statistique de classement :

- Quand on classe une statistique comme statistique de base, est-ce que l'on trouve facilement toutes les statistiques dérivables ?

- Est-ce que classer une statistique de base comme statistique indéterminée (déclasser une statistique de base), pose des difficultés pour trouver les statistiques qui restent dérivables ?

Revenons à notre petit exemple, *Figure 3.1. : Illustration des descriptions statistiques*, repris ci-dessous :

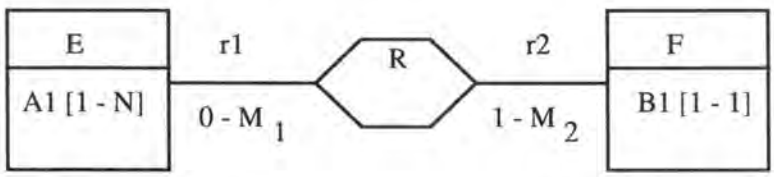


***Figure 4.1.** : Exemple de construction d'un modèle statistique*

Etablissons le système d'équations (d'application), pour plus de facilité dans la construction du modèle statistique :

Partie de gauche du schéma		**Partie de droite du schéma**	
$N_R = N_E \times \mu_{r1}$	Eq1	$N_R = N_F \times \mu_{r2}$	Eq8
$N_E \times \mu_{A1/E} = N_{A1} \times \mu_{E/A1}$	Eq2	$N_F \times \mu_{B1/F} = N_{B1} \times \mu_{F/B1}$	Eq9
$\lambda'_{A1} = \lambda_{A1} / (1 - \Pi o_{A1/E})$	Eq3	$\lambda'_{B1} = \lambda_{B1} / (1 - \Pi o_{B1/E})$	Eq10
$\mu'_{A1/E} = \mu_{A1/E} / (1 - \Pi o_{A1/E})$	Eq4	$\mu'_{B1/E} = \mu_{B1/E} / (1 - \Pi o_{B1/E})$	Eq11
$\mu'_{r1} = \mu_{r1} / (1 - \Pi o_{r1})$	Eq5	$\mu'_{r2} = \mu_{r2} / (1 - \Pi o_{r2})$	Eq12
G1 a un composant A1		B1 composant G2 identifiant	
$\Rightarrow \mu_{G1} = \mu_{E/A1}$	Eq6	$\Rightarrow N_{B1} = N_{G2}$	Eq13
$\Rightarrow N_{A1} = N_{G1}$	Eq7	$\Rightarrow \mu_{G2} = \mu_{F/B1}$	Eq14

***Table 4.1.** : Le système d'équations de l'exemple*

Suivons l'utilisateur dans la construction de son modèle statistique de classement

- Les statistiques suivantes, numérotées de (1) à (7), sont classées comme statistiques dérivées grâce aux contraintes structurelles :

  . $m_{A1} > 0 \quad \Rightarrow \Pi o_{A1/E} = 0$ (1)

. $m_{r2} > 0$ $\Rightarrow \Pi o_{r2} = 0$ (2)

. $m_{B1} = M_{B1} = 1$ $\Rightarrow \mu_{B1/F} = 1$ (3)

. $m_{B1} > 0$ $\Rightarrow \Pi o_{B1/F} = 0$ (4)

. La statistique $\mu'_{B1/F}$ (5) est dérivable

$$\mu'_{B1/F} = \mu_{B1/F} / (1 - \Pi\ o_{B1/F}) \quad (5)$$

. G2 identifiant $\Rightarrow \mu_{G2} = 1$ (6)

. G2 identifiant et a un seul composant $\Rightarrow \mu_{F/B1} = 1$ (7)

- La statistique μ_{r2} (8) est classée comme statistique de base par l'utilisateur, la statistique μ'_{r2} (9) est dérivable : $\mu'_{r2} = \mu_{r2} / (1 - \Pi\ o_{r2})$.
- La statistique μ_{r1} (10) est classée comme statistique de base par l'utilisateur, rien n'est dérivable
- La statistique Πo_{r1} (11) est classée comme statistique de base par l'utilisateur, la statistique μ'_{r1} (12) est dérivable : $\mu'_{r1} = \mu_{r1} / (1 - \Pi\ o_{r1})$.
- La statistique μ_{G1} (13) est classée comme statistique de base par l'utilisateur, la statistique $\mu_{E/A1}$ (14) est dérivable : *G a un composant A* $\Rightarrow \mu_{E/A1} = \mu_{G1}$
- La statistique N_{A1} (15) est classée comme statistique de base par l'utilisateur, la statistique N_{G1} (16) est dérivable : *G a un composant A* $\Rightarrow N_{G1} = N_{A1}$
- La statistique $\mu_{A1/E}$ (17) est classée comme statistique de base par l'utilisateur, les statistiques de (18) à (23) sont dérivées grâce aux équations suivantes :

  (18). $N_E = (N_{A1} \times \mu_{E/A1}) / \mu_{A1/E}$

  (19). $\mu'_{A1/E} = \mu_{A1/E} / (1 - \Pi o_{A1/E})$

  (20). $N_R = N_E \times \mu_{r1}$

  (21). $N_F = N_R / \mu_{r2}$

  (22). $N_{B1} = (N_F \times \mu_{B1/F}) / \mu_{F/B1}$

  (23). $N_{G2} = N_{B1}$
- La statistique λ_{A1} (24) est classée comme statistique de base par l'utilisateur, la statistique λ'_{A1} (25) est dérivable : $\lambda'_{A1} = \lambda_{A1} / (1 - \Pi o_{A1/E})$.
- La statistique λ'_{B1} (26) est classée comme statistique de base par l'utilisateur, la statistique λ_{B1} (27) est dérivable : $\lambda_{B1} = \lambda'_{B1} \times (1 - \Pi o_{B1/F})$.

Résumons le modèle statistique de classement ainsi construit par l'utilisateur dans une table. Les statistiques sont suivies de leur numéro d'ordre de classement et leur classe.

Partie de gauche du schéma			Partie de droite du schéma		
$\Pi o_{A1/E} = 0$	(1)	D	Πo_{r2}	(2)	D
μ_{r1}	(10)	B	$\mu_{B1/F}$	(3)	D
Πo_{r1}	(11)	B	$\Pi o_{B1/F}$	(4)	D
μ'_{r1}	(12)	D	$\mu'_{B1/F}$	(5)	D
μ_{G1}	(13)	B	μ_{G2}	(6)	D
$\mu_{E/A1}$	(14)	D	$\mu_{F/B1}$	(7)	D
N_{A1}	(15)	B	μ_{r2}	(8)	B
N_{G1}	(16)	D	μ'_{r2}	(9)	D
$\mu_{A1/E}$	(17)	B	N_F	(21)	D
N_E	(18)	D	N_{G2}	(22)	D
$\Pi o_{A1/E}$	(19)	D	N_{B1}	(23)	D
N_R	(20)	D	λ'_{B1}	(26)	B
λ_{A1}	(24)	B	λ_{B1}	(27)	D
λ'_{A1}	(25)	D			

***Table 4.2.** : Modèle statistique de classement de l'exemple*

Essayons de répondre à nos deux questions :

4.2.1. Classer une statistique indéterminée comme statistique de base

Quand on classe une statistique indéterminée comme statistique de base, est-ce que toutes les statistiques dérivables ont été obtenues ? C'est difficile de l'affirmer. L'exemple comme il est présenté ne laisse pas entrevoir les difficultés qui peuvent se produire.

4.2.2. Déclasser une statistique de base

Déclasser une statistique de base revient à classer une statistique de base comme statistique indéterminée.

Reprenons notre exemple, *Figure 4.1. : Exemple de construction d'un modèle statistique* et posons-nous la question pour une statistique de base. Si on remet la statistique μ_{r1} à indéterminée, quelles statistiques ne sont plus dérivables ?

Regardons ce qu'il en est des statistiques qui peuvent être directement dérivées à partir de μ_{r1} : N_R (Dérivée), N_E (Dérivée), μ'_{r1} (Dérivée), Πo_{r1} (de Base). Trois d'entre elles sont dérivées. Elles sont remises en question ! En plus il faut remettre en questions toutes les statistiques qui peuvent être directement dérivées à partir de celles-ci (indirectement de μ_{r1}), et ainsi de suite. Bref le modèle statistique de classement en entier est remis en question. Ici aussi l'exemple comme il est présenté ne permet pas de l'affirmer.

4.3. Analyse et solutions des problèmes

4.3.1. Expression de l'exemple sous forme d'un graphe d'équations

Une représentation graphique des équations de cet exemple nous permet de le commenter plus aisément dans la suite. Ce graphe montre les statistiques qui interviennent dans les équations. Une équation est représentée par son numéro dans un cercle. La participation d'une statistique à l'équation est indiquée par une arête les reliant.

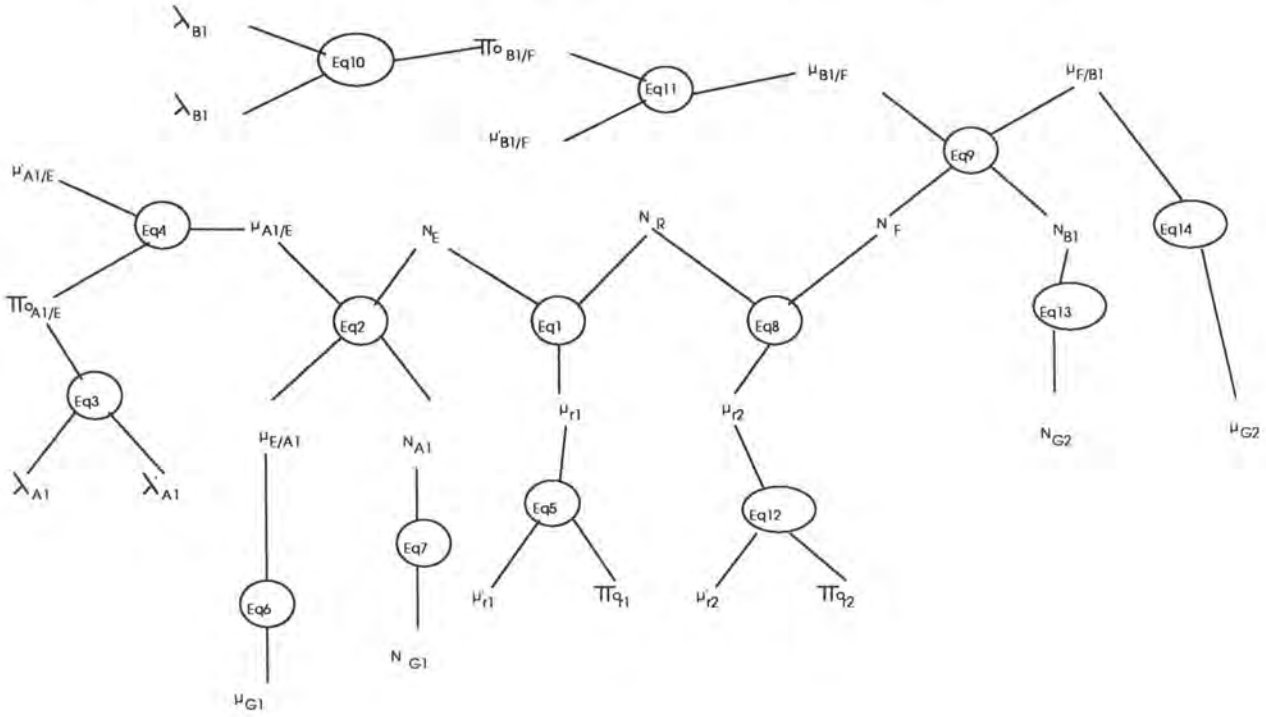


Figure 4.2. : *Graphe d'équations d'un modèle statistique*

4.3.2. Classer une statistique comme statistique de base

Nous allons montrer et analyser le problème qui peut se poser lorsqu'on classe une statistique comme statistique de base. Le problème se produit quand une statistique dérivable peut être dérivée de plusieurs façons. Nous avons de l'information redondante.

Supposons que l'on rajoute une contrainte d'égalité entre groupes au schéma. Le système d'équations est augmenté de deux équations. Posons-nous la question suivante :

Est-ce que la contrainte d'égalité entre groupes pose un problème ?

Contrainte d'égalité entre groupes :

$\Rightarrow N_{G1} = N_{G2}$ Eq15

$\Rightarrow \mu_{G1} = \mu_{G2}$ Eq16

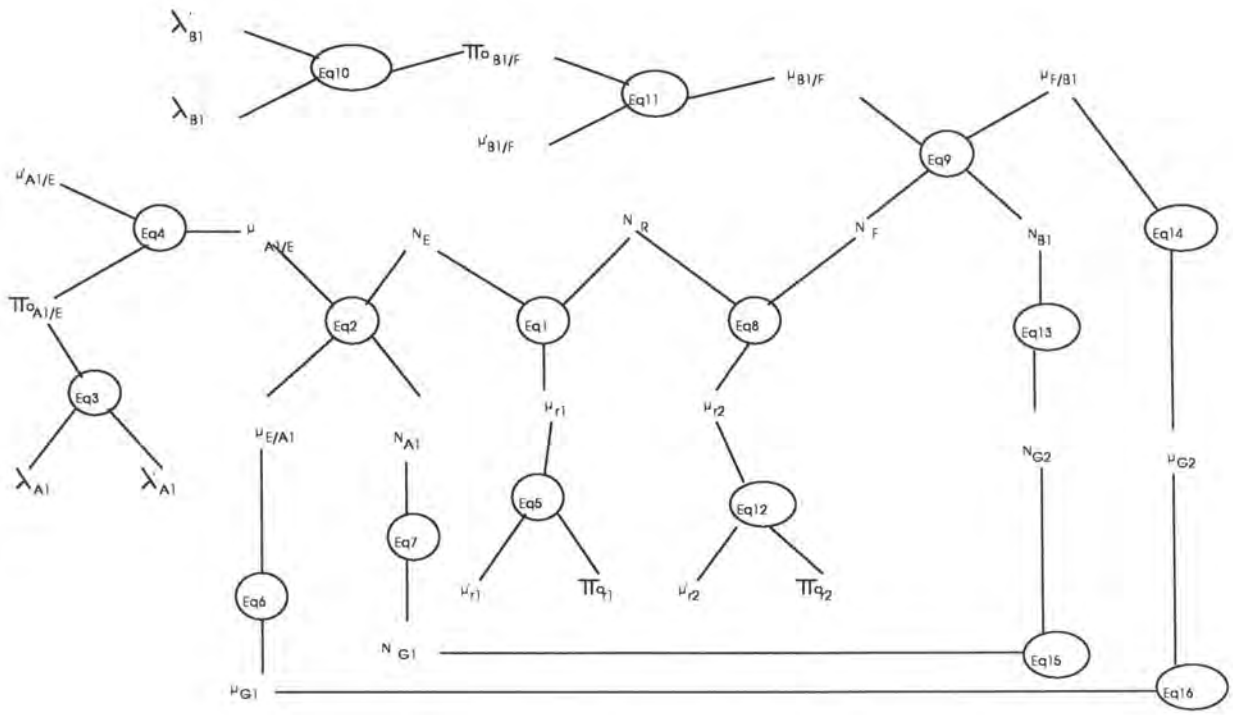


***Figure 4.3.** : Un modèle statistique avec une contrainte d'égalité*

Nous apercevons tout de suite qu'il s'est créé des circuits. Enumérons la suite d'équations qui les composent :

1 circuit : 1, 2, 6, 16, 14, 9,8 et 1.
2 circuit : 1, 2, 7, 15, 13, 9, et 8.
3 circuit : 2, 6, 16, 14,9 13, 15, 7, 2.

La présence d'un circuit est importante puisqu'elle offre la possibilité de construction d'un mauvais modèle statistique, car nous pouvons dériver une statistique de plusieurs façons. Considérons le circuit 3 et simplifions pour nous en rendre clairement compte :

$\mu_{E/A1} = \mu_{G1}$ et $\mu_{G1} = \mu_{G2}$ et $\mu_{F/B1} = \mu_{G2}$ $\Rightarrow \mu_{F/B1} = \mu_{E/A1}$

$N_{A1} = N_{G1}$ et $N_{G1} = N_{G2}$ et $N_{B1} = N_{G2}$ $\Rightarrow N_{B1} = N_{A1}$

Le circuit peut donc être écrit sous la forme suivante

$N_E \times \mu_{A1/E} = N_{A1} \times \mu_{E/A1}$ Eq2
$N_F \times \mu_{B1/F} = N_{A1} \times \mu_{E/A1}$ Eq17

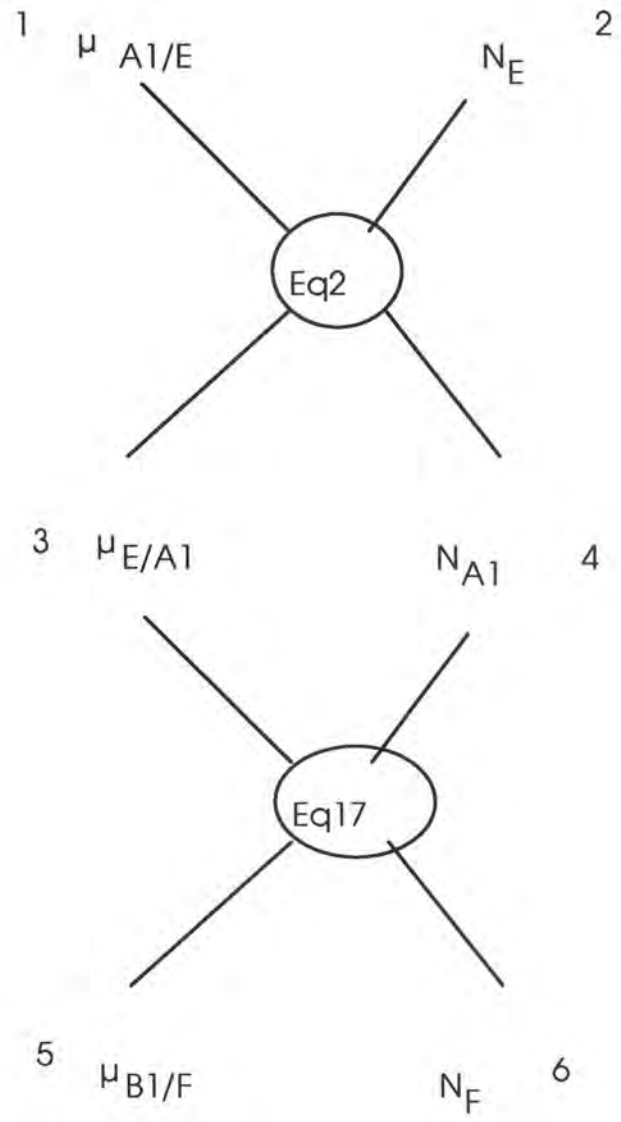


***Figure 4.4.** : Graphe d'équations d'un circuit réduit*

Supposons que la construction du modèle statistique de classement commence avec le classement des statistiques 1, 2, 5 et 6 comme statistiques de base. Rien n'est encore dérivable à partir de cet ensemble. Toutefois la prochaine statistique classée comme statistique de base, statistique 3 par exemple, qui permet de dériver la statistique 4 de deux façons : à partir des statistiques 1, 2 et 3 ainsi qu'à partir des statistiques 3,5 et 6. Nous allons analyser une situation comme celle-ci en introduisant le concept de modèle statistique de classement redondant qui exprime la situation différemment.

Définition d'un modèle statistique de classement redondant

> Un modèle statistique de classement redondant comprend au moins une statistique de base qui est dérivable à partir d'autres statistiques de base.

Exprimons la situation de l'exemple en ces termes. Le classement de la statistique 4 comme de base rend le modèle statistique de classement redondant : à partir du sous-ensemble de statistiques de base 1, 2, 3 et 5 on peut dériver la statistique de base 6. Ceci en dérivant la statistique 4 à partir des statistiques 1,2 et 3; et ensuite la statistique 6 à partir des statistiques 3, 4 et 5.

Est-ce que les équations sont sous la bonne forme ? Remplaçons le système d'équations par un système d'équations équivalent (l'équation Eq18 est la combinaison des deux autres) :

$$\begin{cases} N_E \times \mu_{A1/E} = N_{A1} \times \mu_{E/A1} & \text{Eq2} \\ N_E \times \mu_{A1/E} = N_F \times \mu_{B1/F} & \text{Eq18} \end{cases}$$

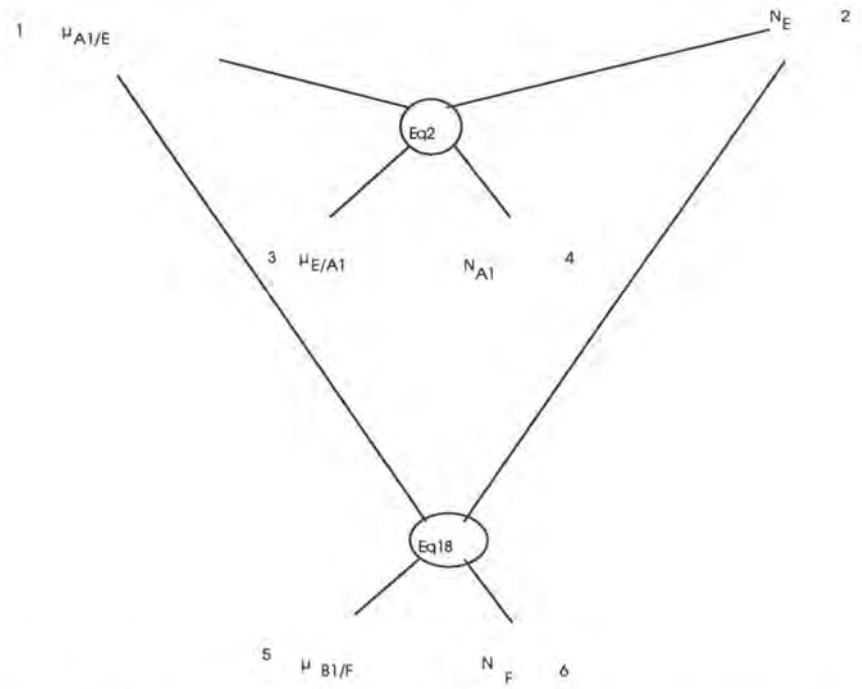


***Figure 4.5.** : Graphe d'équations équivalent du circuit réduit*

Le graphe prend une autre forme, pour lequel le problème ci-dessous ne se produira pas. Lors de la construction la statistique 4 ne peut plus être classée comme statistique de base parce qu'elle est dérivée des statistiques 1,2 et 3. Néanmoins le problème ne se trouve pas encore résolu puisque même sous cette forme il y a la possibilité de construire un modèle redondant. Par exemple le classement des statistiques 3,4, 5 et 6 comme statistiques de base nous y entraîne.

On peut se rendre compte d'une manière différente qu'il y a un problème. Considérons un modèle statistique avec un nombre V de statistiques, dont un nombre B sont des statistiques de base, et dans un système avec un nombre E d'équations. Le modèle est redondant si [B > V- E] ou si [B = V-E et ne sont pas encore toutes les statistiques dérivables].

Revenons à notre exemple Figure 4.1. : Exemple de construction d'un modèle statistique. L'exemple vérifie l'équation B = V-E : 8 statistiques de base = 27 variables - 14 équations - 5 statistiques connues grâce aux contraintes structurelles. La non-redondance de ce modèle est garantie.

4.3.3. Construire des modèles statistique non-redondants

Maintenant que nous avons un exemple d'un problème, nous allons analyser s'il y en a d'autres et comment nous pouvons les résoudre.

Analysons s'il a d'autres possibilités de créer un circuit. Pour cela nous devons nous pencher sur les types d'équations que nous avons présentés dans le point *3.1.2. Modèle statique des statistiques*. Les équations qui résultent dans des constantes ne posent pas de problèmes. Les équations sous contraintes structurelles d'un groupe ont été vérifié avec l'exemple, *Figure 4.3. : Un modèle statistique avec une contrainte d'égalité* , et nous avons découvert que la contrainte d'égalité entre groupe pose un problème. Les équations référencées T1 jusqu'à T5 ne posent pas de problèmes, puisque aucun circuit ne peut être créé. Il nous reste donc deux équations, T6 et T7, à vérifier :

Est-ce que la relation récursive pose un problème ?

Examinons cette question avec l'exemple suivant :

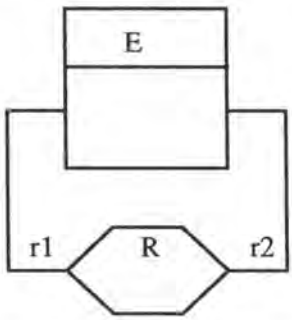


***Figure 4.6.** : La relation récursive*

A partir des types d'équations T1 et T6 (nous n'examinons que celles-ci) nous pouvons établir les équations de ce schéma :

$$
\begin{array}{|ll}
N_R = N_E \times \mu_{r1} & 1 \\
N_R = N_E \times \mu_{r2} & 2 \\
\text{relation récursive} & \\
\quad \Rightarrow \mu_{r1} = \mu_{r2} & 3
\end{array}
$$

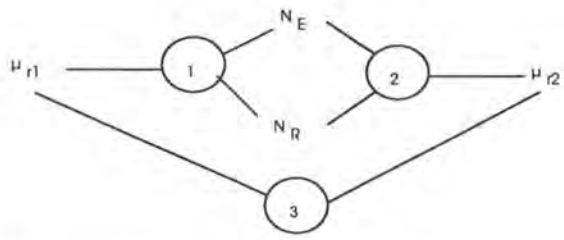


Figure 4.7. : *Graphe d'équations de la relation récursive*

Il y a des circuits, toutefois une construction d'un modèle statistique de classement redondant est impossible. La cause du circuit réside dans la redondance du système d'équations, et de ce fait le circuit n'existe pas vraiment.

Définition d'un système d'équations redondant

> Un système d'équations est redondant quand une de ses équations peut être éliminée parce qu'elle est une combinaison des autres équations.

Réduire cette redondance dans le système d'équation n'est pas une solution, puisque alors certains modèles statistique ne sont plus possibles.

Est-ce que l'attribut décomposable pose un problème ?

Analysons cela avec un exemple :

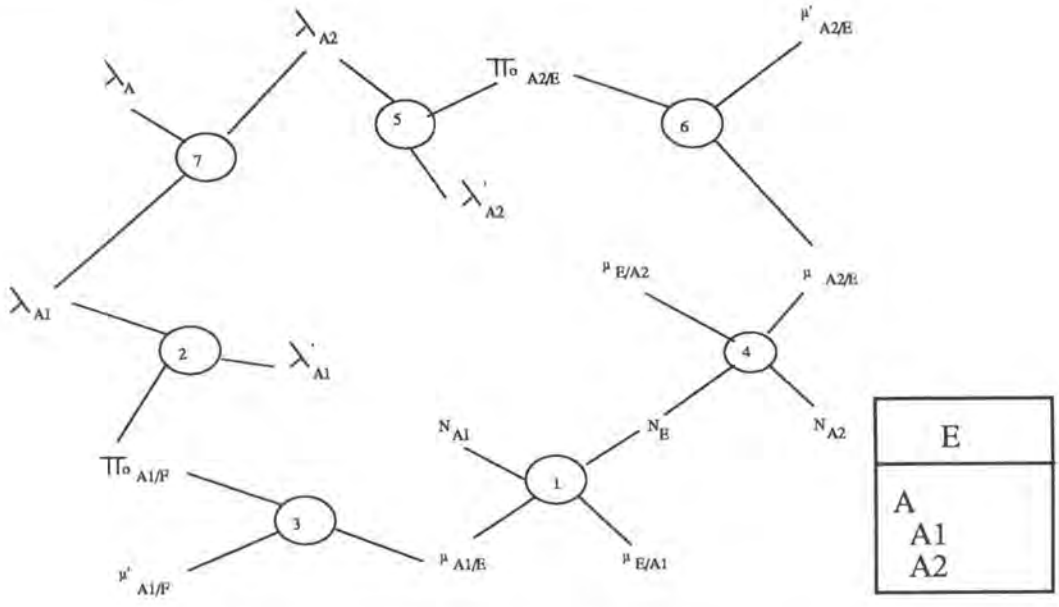


Figure 4.8. : *Un attribut décomposable avec son graphe d'équations*

Nous avons le système d'équations suivant :

$N_E \times \mu_{A1/E} = N_{A1} \times \mu_{E/A1}$ 1
$\lambda'_{A1} = \lambda_{A1} / (1 - \Pi o_{A1/E})$ 2
$\mu'_{A1/E} = \mu_{A1/E} / (1 - \Pi o_{A1/E})$ 3
$N_E \times \mu_{A2/E} = N_{A2} \times \mu_{E/A2}$ 4
$\lambda'_{A2} = \lambda_{A2} / (1 - \Pi o_{A2/E})$ 5
$\mu'_{A2/E} = \mu_{A2/E} / (1 - \Pi o_{A2/E})$ 6
$\lambda_A = \lambda_{A1} + \lambda_{A2}$ 7

Avec le graphe d'équations, nous pouvons affirmer qu'un attribut ne pose pas de problème. Donc finalement, c'est uniquement la contrainte d'égalité entre groupe qui peut créer un problème.

Comment construire un modèle statistique non-redondant avec les pièges introduit par les contraintes d'égalité entre groupes ? Pour la gestion, trois solutions viennent à l'esprit :

1. Enlever les pièges
2. Utiliser un solveur intelligent
3. Déceler et éviter les pièges

a) Enlever les pièges

Si nous prenons les deux graphes en même temps (*Figure 4.4. : Graphe d'équations d'un circuit réduit* et *Figure 4.5. : Graphe d'équations équivalent du circuit réduit*), la construction d'un modèle statistique redondant est impossible : il n'y a plus de piège.

Le piège constitué de l'équation Eq2 et Eq17

$$\begin{cases} N_E \times \mu_{A1/E} = N_{A1} \times \mu_{E/A1} & \text{Eq2} \\ N_{A1} \times \mu_{E/A1} = N_F \times \mu_{B1/F} & \text{Eq17} \end{cases}$$

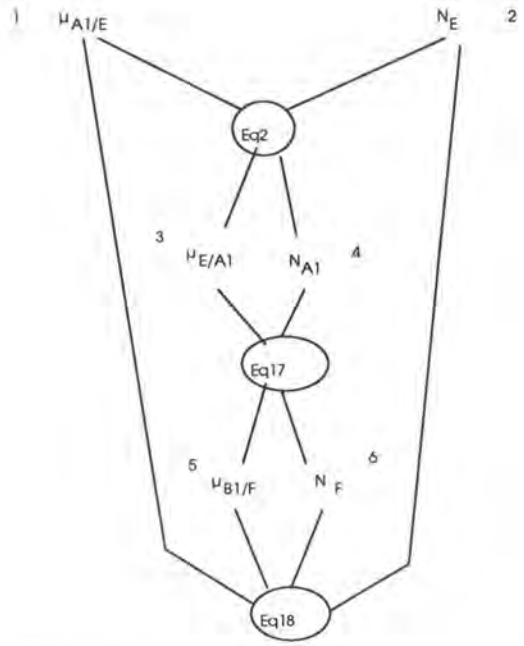


Figure 4.9. : *Graphe d'équations redondant du circuit réduit*

La solution consiste donc à ajouter l'équation $N_E \times \mu_{A1/E} = N_F \times \mu_{B1/F}$ Eq18 au système d'équations.

Cette première solution pour construire un modèle statistique de classement non-redondant est facile à réaliser. Rajouter une équation, c'est enlever le piège et l'utilisateur est à l'abri d'une mauvaise construction. Cette solution est (aussi) d'application dans notre problème initial (Figure 4.3. : Un modèle statistique avec une contrainte d'égalité) : à l'introduction de la contrainte de référence dans notre exemple.

b) Un solveur intelligent

Cette solution ainsi que la suivante consiste à déceler le piège à la construction d'un modèle statistique de classement par l'utilisateur. Un solveur intelligent choisit le système d'équation qui convient (pour que le modèle reste non-redondant).

Déceler un piège, revient à rechercher un (les) circuit(s) (créés par des contraintes d'égalité de groupes) d'un graphe. Pour un tel algorithme nous renvoyons à la littérature [FICH].

Rappelons que l'on se sert aussi d'un de ces algorithmes dans le point suivant.

c) Déceler et éviter les pièges

Une des statistiques de base du piège n'est pas considérée comme de base. Sa valeur devra être validée par sa valeur dérivée.

Cette manière de faire offre la possibilité d'ouvrir la gestion à des modèles a priori redondants en bénéficiant toujours des avantages des modèles non-redondants. Par exemple supposons qu'une gestion de statistiques impose à l'utilisateur les populations des type d'entité comme statistiques de base. La construction peut aboutir à des modèles redondants. Un exemple évident dans ce cas, où une des deux valeurs suffit :

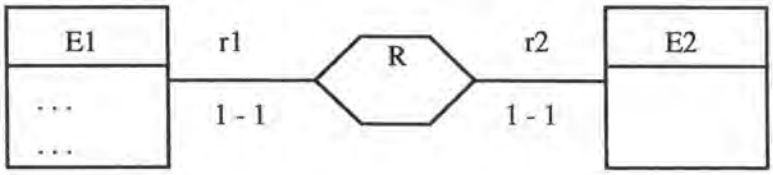


***Figure 4.10.** : Exemple d'un modèle statistique a priori redondant*

Des gestions de statistiques peuvent aboutir à des idées pour guider un utilisateur à travers la construction du modèle statistique. Des propositions de gestion de ce genre dépassent le cadre de ce mémoire. Le lecteur intéressé dans une telle gestion se référera à l'article [HAI,92].

4.3.4. Expression de l'exemple sous forme d'un graphe des dépendances

Lors de la construction du modèle statistique de classement de notre exemple *Figure 4.2. : Exemple de construction d'un modèle statistique,* le sens de dérivation est fixé : chaque équation permet de dériver une statistique. Cette statistique dépend des statistiques à partir desquelles elle est dérivée. La représentation graphique des dépendances entre les statistiques, nous permet de commenter l'exemple plus aisément dans la suite.

Les arcs représentent la participation d'une statistique (début de l'arc, ascendant) à la dérivation d'une statistique dérivée (fin de l'arc, descendant).

Les niveaux de dérivation correspondent au nombre de dérivations qu'il faut faire pour dériver la statistique.

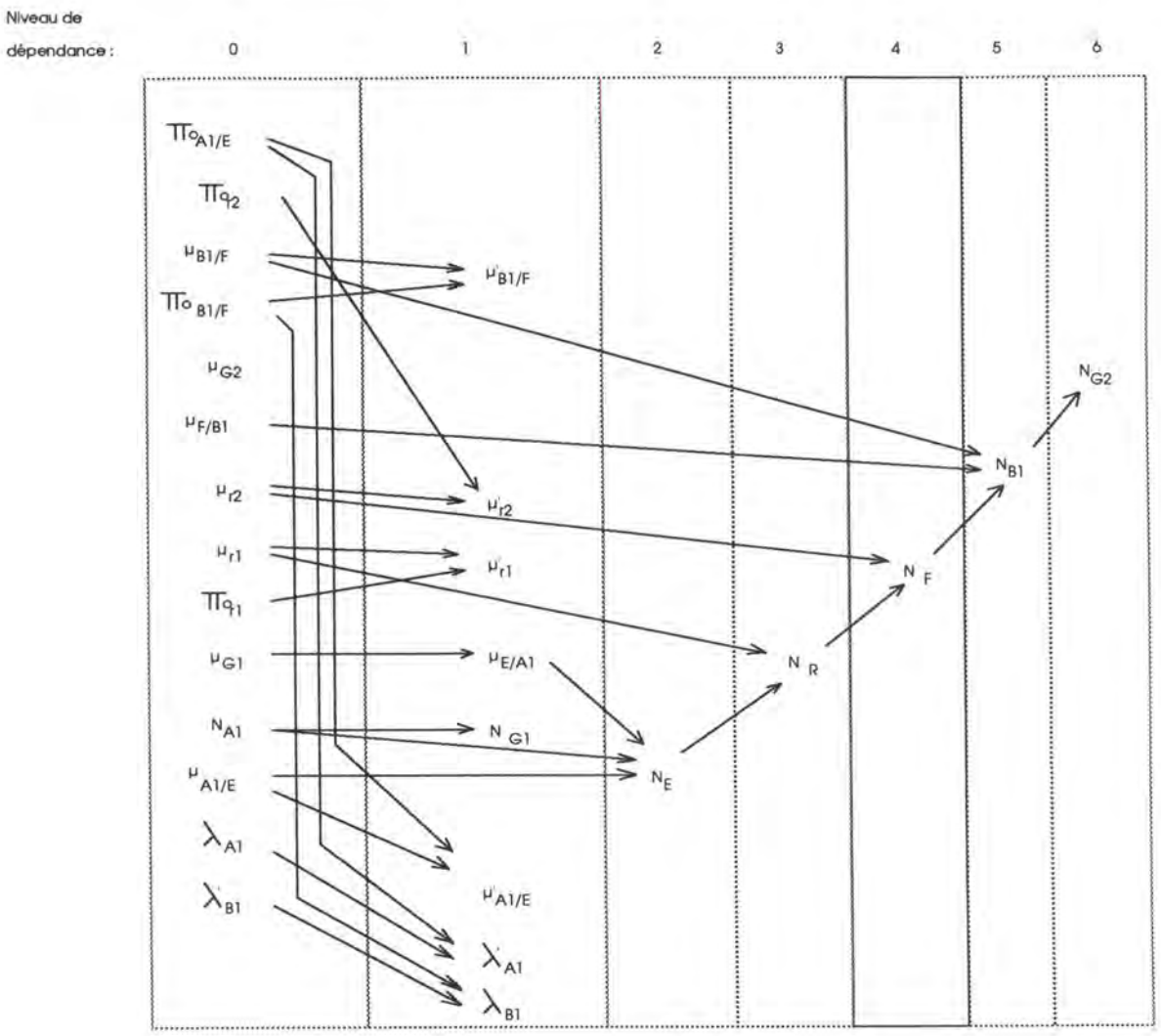


Figure 4.11. : *Exemple d'un graphe des dépendances*

Le graphe des dépendances permet facilement de trouver les statistiques dérivables qui dépendent d'une statistique de base (les descendants) et de trouver de quelles statistiques de base une statistique dérivée dépend (les ascendants).

Ainsi , nous pouvons dire que N_R dépend directement de μ_{r1} et indirectement aussi de N_{A1}, $\mu_{A1/E}$ et de μ_{G1}. C'est l'ensemble des ascendants du niveau de dépendance 0.

Ainsi , les statistiques qui dépendent de μ_{r2} sont les statistiques dérivées μ'_{r2} ,N_F et N_{B1}.

4.3.5. Déclasser une statistique de base

Le graphe des dépendances nous permet de facilement trouver ce qui n'est plus dérivable quand on déclasse une statistique de base. Ce sont tous les descendants.

Si nous déclassons par exemple μ_{r1}, les statistiques μ'_{r1}, N_R, N_F, N_{B1} et N_{G2} ne sont plus dérivables.

4.3.6. Calcul d'un modèle statistique

a) Calcul d'un modèle statistique non-redondant

L'avantage des modèles statistiques non-redondants réside dans la validation des valeurs. Celle-ci se fait par rapport aux contraintes structurelles. Une statistique de base en plus est inutile et apporte du travail : elle doit être validée par rapport aux contraintes et par rapport à l'ensemble des statistiques non-redondants. Si la statistique de base inutile n'est pas valide. Toutes les statistiques dérivées de celle-ci ne le sont pas non plus.

Reprenons un exemple, *Figure 4.4. : Graphe d'équations d'un circuit réduit*, pour faciliter la compréhension de ce qui est dit : , et supposons la construction du modèle statistique de classement redondant, constitué des statistiques de base 1, 2, 5, 6, 3. (Remarquons qu'ils sont énumérées dans ordre de la construction qui aboutit à un modèle statistique de classement redondant). La statistique 4 est dérivée.

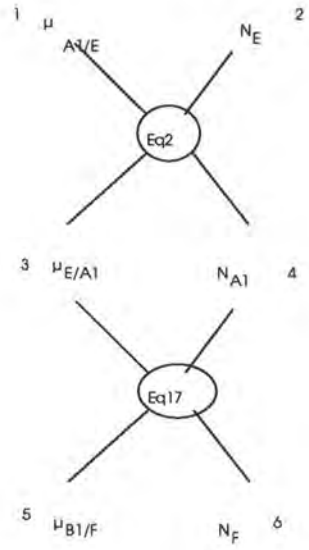


Figure 4.12. : *Statistiques dérivées dans un modèle statistique de classement redondant*

Sur cet exemple, nous pouvons reprendre notre exemple. La statistique 4 peut être dérivée de deux ensemble de statistiques de base différent : l'ensemble 1, 2, 3 et l'ensemble 3, 5, 6. Supposons que la statistique 5 n'est pas valide, et que les statistiques 1, 2, 3 sont valides. La non-validité de la statistique 4 peut alors être découverte par dérivation de l'ensemble des statistiques 1, 2 et 3, et implique la non-validité des statistiques dérivées de celle-ci.

b) Calcul d'un modèle statistique acceptable

Le calcul d'un modèle statistique de valeurs à partir du modèle statistique de classement n'apporte pas toujours une solution acceptable pour le concepteur. Trois cas peuvent se produire :

- Un ensemble de valeurs nulles, lequel n'est pas intéressant pour le concepteur.

- Le modèle statistique de classement ne possède aucun modèle statistique de valeurs. Ce cas ci se produit quand l'ensemble des contraintes structurelles est contradictoire. La validité d'un schéma est étudiée dans le Chapitre III Règles de validation des spécifications de la référence [BOD-PIGN, 83], mais ne commente pas la validité de l'ensemble des contraintes d'un schéma (ou consistance). La validation de la consistance d'un schéma est encore du domaine de recherche (car c'est un problème fort complexe) à ma connaissance et sort du cadre du mémoire.

  Donnons un exemple pour concrétiser le problème.

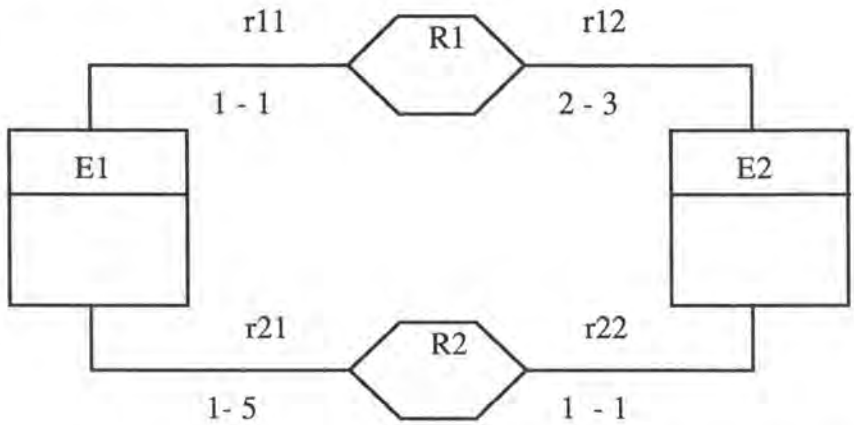


Figure 4.13. : *Un ensemble de contraintes structurelles non-valide*

Une base de données, décrit par un tel schéma, doit respecter le système d'équations suivant :

$$\begin{cases} N_{R1} = N_{E1} \times \mu_{r11} \\ N_{R1} = N_{E2} \times \mu_{r12} \\ N_{R2} = N_{E1} \times \mu_{r21} \\ N_{R2} = N_{E2} \times \mu_{r22} \end{cases}$$

Transformons le système d'équations dans le suivant :

$$\begin{cases} N_{E1} \times \mu_{r11} = N_{E2} \times \mu_{r12} \\ N_{E1} \times \mu_{r21} = N_{E2} \times \mu_{r22} \end{cases}$$

Examinons ce système d'équations en plus près. Deux équations, et dans les deux il y a N_{E1} et N_{E2}. Posons-nous la question : Quelle solution y a t'il pour N_{E1} et N_{E2} ?

Le système d'équations n'a pas une solution acceptable puisque $\mu_{r11} \times \mu_{r22} \neq \mu_{r21} \times \mu_{r12}$ et la seul solution possible est que N_{E1}, N_{E2} , N_{R1}et N_{R1} soient égal à zéro.

- Lors de la dérivation de la valeur d'une statistique dérivée, une contrainte structurelle est violée. Une question se pose alors : est-ce que nous pouvons trouver les statistiques de base qui sont à l'origine de cette valeur non-valide ?

Exemple

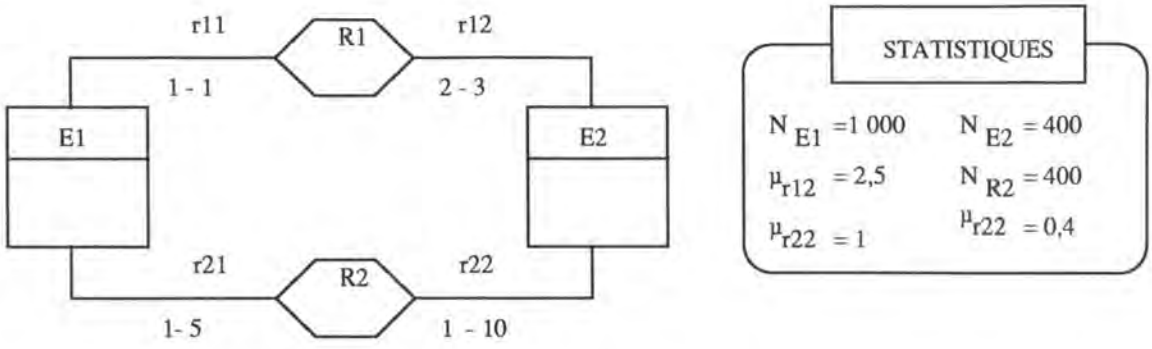


Figure 4.14. : *Un modèle statistique des valeurs non-valide*

Supposons la construction du modèle statistique suivant :

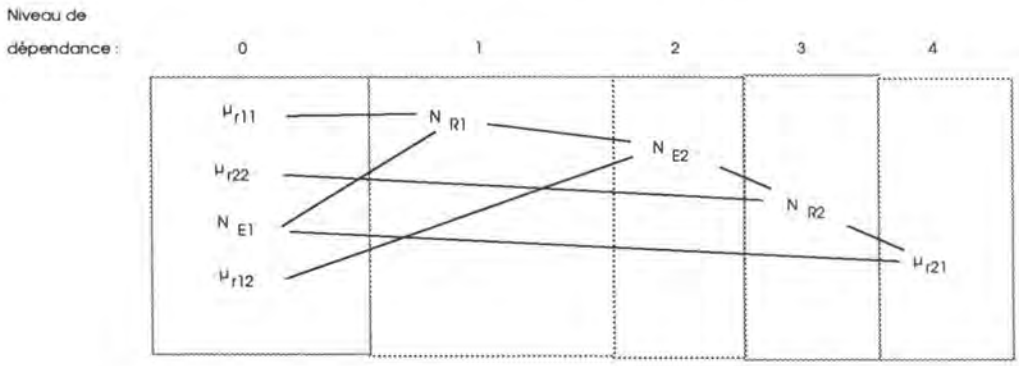


Figure 4.15. : *Un modèle statistique des valeurs non-valide*

Supposons que le concepteur introduit $N_{E1} = 1000$, $\mu_{r12} = 2,5$ et $\mu_{r22} = 1$. Ces deux statistiques de base respectent les contraintes structurelles.

$N_{R1} = N_{E1} = 1000$
$N_{E2} = N_{R1} / \mu_{r12} = 1000 / 2,5 = 400$
$N_{R2} = N_{E2} \times \mu_{r22} = 400 \times 1 = 400$
$\mu_{r21} = N_{R2} / N_{E1} = 0,4$ la contrainte structurelle est violée !

Une correction de la valeur d'une des statistiques ascendants de base peut faire l'affaire, soit N_{E1}, μ_{r12} ou μ_{r22}. Prenons pour $\mu_{r22} = 5$, et on obtient $\mu_{r21} = 2$ et la contrainte structurelle est respectée, cette fois.

La gestion va permettre de changer les valeurs des statistiques de base qui sont à l'origine de cette valeur non-valide, pour faire évoluer le modèle statistique des valeurs vers un modèle statistique des valeurs valides. Le graphe des dépendances de la *Figure 4.11. : Exemple d'un graphe des dépendances* , montre bien qu'une statistique dépend de tous ses ascendant du niveau de dépendance 0.

4.3.7. Transformation du modèle statistique

La transformation d'un schéma dans un autre, nécessite la transformation du modèle statistique. Dans le nouveau schéma les statistiques déduites de la structure et les statistiques transformées de l'ancien n'existent plus, disons que la **destruction** de l'ancienne structure les détruit aussi. D'autres statistiques expriment la nouvelle structure, disons que la **création** de la nouvelle structure les créé.

La destruction et la création de statistiques change le modèle statistique. Pour la création, il faut prendre en compte l'état d'évolution du modèle. Analysons les deux cas de figures suivants, dans lesquelles la transformation du modèle statistique réagit de façon différente :

1. Le modèle statistique de classement est construit.

La transformation consiste à (r)établir le classement des nouvelles statistiques. Les équations de transformations, établies dans le point *3.3. Transformations des statistiques*, conviennent pour cela. Celles qui ont plus d'un argument résultent dans la classe de statistique de base si tous les arguments sont des statistiques de base sinon il faut vérifier les équations que la statistique joue dans le schéma.

Considérons la transformation 3.3.2.A. *Transformation d'un attribut en type d'entité par représentation par valeurs*. Une équation de transformation à plusieurs arguments est par exemple : $N_R = N_E \times \mu_{A/E}$, (arguments du premier schéma) dont un des deux arguments est dérivée. Toutefois la statistique N_R est une statistique de base quand dans l'équation $N_R = N_{E2} \times \mu_{r2}$ (arguments du premier schéma) les deux statistiques sont classées comme statistiques de base.

La transformation ne s'occupe pas des statistiques indéterminées.

2. Le modèle statistique des valeurs valides est établi.

Dans ce cas-ci les équations de transformations forment un outil complet pour réaliser la transformation d'un modèle statistique avec ses valeurs.

4.4. Proposition de gestion

4.4.1. Description de la solution

Résumons encore brièvement le problème, de la gestion (souple) des statistiques dans la conception de base de données. Les statistiques d'un schéma forment un système d'équations dans lequel il y a plus de variables que d'équations. La construction d'un modèle statistique consiste à fixer certaines de ses statistiques en introduisant des valeurs (fixer des variables comme constantes) et au fur et à mesure le système d'équations devient soluble (quand le nombre de variables équivaut au nombre d'équations). Toutefois on s'aperçoit que, tout ensemble de variables fixées comme constantes ne convient pas (le nombre d'équations équivaut au nombre de variables et le système d'équations n'est pas encore résolu).

Définition d'un modèle statistique de classement redondant

Un modèle statistique de classement redondant comprend une statistique de base qui est dérivable à partir d'autres statistiques de base.

Définition d'un modèle statistique des valeurs valides

Un modèle statistique des valeurs est valide si les statistiques de base ainsi que les statistiques dérivées respectent les contraintes structurelles.

Définition d'un schéma valide

Un schéma valide possède au moins un modèle statistique des valeurs valides.
(Un schéma qui n'est pas valide, ne peut posséder une extension. C'est un schéma purement conceptuel)

Enonçons quelques propriétés. Nous en avons besoin pour proposer l'algorithme de construction d'un modèle statique non-redondant et l'algorithme de calcul d'un modèle statistique des valeurs.

1.

Si on ajoute au système d'équations composé des 7 types d'équations énumérés dans le point *3.1.2. Modèle statique des statistiques*, le type d'équation sous contrainte structurelle suivant :

Contrainte de référence entre deux groupes constitués d'un composant

$$\Rightarrow N_E \times \mu_{A1/E} = N_F \times \mu_{B1/F}$$

Alors la construction d'un modèle statistique de classement à partir de ces types d'équations aboutit à un modèle statistique de classement non-redondant.

Cette propriété a été analysé et résolu dans le point 4.3.3. Construire des modèles statistique non-redondants. Elle permet de construire des modèle statistique non-redondant.

2.

Les valeurs des statistiques dérivées d'un modèle statistique des valeurs, obtenue à partir d'un modèle statistique des valeurs de base non-redondant, respectent toutes les équations auxquelles elles participent.

Pour s'en convaincre il suffit de supposer le contraire : les deux valeurs doivent être dérivées de deux ensembles de statistiques de base différentes, ce qui est absurde pour un modèle statistique non-redondant.

Cette propriété est utile pour l'algorithme de calcul : les valeurs ne doivent pas être vérifiées les unes par rapport aux autres, uniquement la correction des contraintes structurelles doit être vérifiée.

3.

Si l'ensemble des contraintes d'un schéma est cohérent (sans contradiction), alors le schéma est valide et il existe au moins un modèle statistique des valeurs valides pour ce schéma.

Cette propriété a été analysée dans le point *4.3.6. Calcul d'un modèle statistique.*

4.4.2. Les objets abstraits des algorithmes

Les algorithmes qui suivent utilisent les objets abstraits suivants :

- Séquence des statistiques de base
  Retient dans un ordre chronologique les statistiques classées comme statistiques de base par l'utilisateur.

- Séquence des statistiques dérivées
  Retient dans un ordre chronologique les statistiques dérivées.

- La structure abstraite du modèle statistique
  Cette structure de données est composée de la séquence des statistiques de base, de la séquence des statistiques dérivées, ainsi que des liens entre ces deux séquences. Les liens montrent les statistiques dérivables à partir des statistiques de base ou dérivables qui précèdent. Par exemple la statistique dérivée numéro 4 peut être dérivée à partir de certains (ou de toutes les) statistiques de l'ensemble formé par les statistiques 1 et 2 de la séquence des statistiques de base et les statistiques 1, 2 et 3 de la séquence des statistiques dérivées.

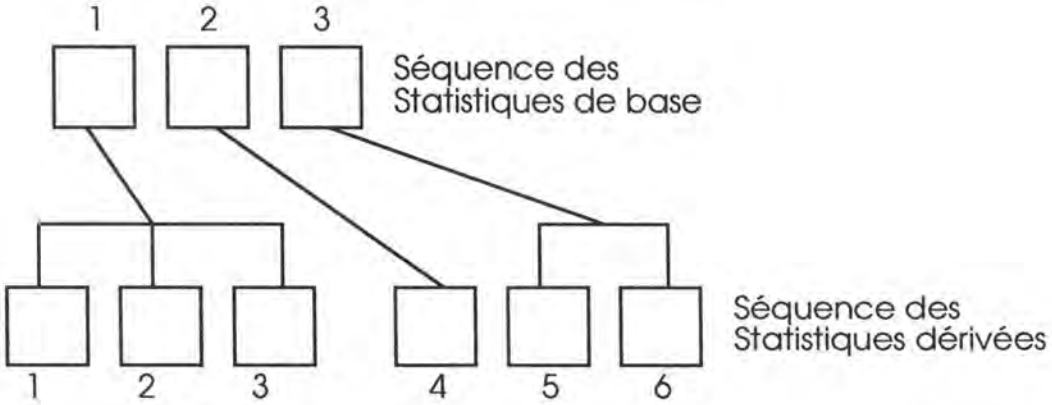


***Figure 4.16. :** Exemple de structure du modèle statistique*

- Les équations de dérivation

L'algorithme de construction d'un modèle statistique de classement à besoin des équations par statistique dans laquelle elle intervient comme argument. Ce sont les équations d'application du point 3.*1.2. a) Les types équations* et du point 3.1.2.. *b) Les constantes et les équations sous contrainte structurelle* mais il faut les réarranger pour les avoir sous la forme désirée.

Pour obtenir par exemple les équations avec N_E comme argument, à partir de l'équation $N_E = N_R / \mu_{ri}$, on doit réarranger l'équation pour chaque argument, et on obtient les deux équations $N_R = N_E \times \mu_{ri}$ et $\mu_{ri} = N_R / N_E$.

Pour les raisons commentées au point *4.3.3. a) Enlever les pièges* et appuyé par le point *4.4. Proposition de gestion* nous rajoutons l'équation suivante T8 dans le cas d'une contrainte de référence entre deux groupes ayant chacun un composant :

$$N_E \times \mu_{A1/E} = N_F \times \mu_{B1/F} \qquad T8$$

4.4.3. Algorithme abstrait de construction d'un modèle statistique de classement non-redondant

Algorithme abstrait, sous forme informel, de construction d'un modèle statistique de classement non-redondant :

1. Les statistiques dérivables à partir des équations sous contraintes structurelles sont classées comme dérivées et ajoutées à la séquence des statistiques dérivées.

2. Classer une statistique comme statistique de base

- On ajoute la statistique à la séquence des statistiques de base;
- On sélectionne toutes les équations ayant cette statistique comme argument et dont les autres arguments sont des statistiques de base ou des statistiques dérivées. Les statistiques dérivables avec ces équations sont rajoutées à la séquence des statistiques dérivées. On réitère ce point pour chaque statistique obtenue (on recherche ainsi les statistiques qui sont indirectement dérivables).

3. Classer une statistique de base comme statistique indéterminée

- On enlève la statistique de la séquence des statistiques de base;
- On garde les statistiques dérivées avant le classement de cette statistique et on remet toutes les statistiques dérivables suivantes à indéterminées. On réitère le point 2 pour les statistiques de base situées après celle-ci.

4.4.4. Algorithme de calcul et de validation des statistiques dérivées d'un modèle statistique de classement non-redondant

La structure du modèle statistique est utilisée. L'ordre de calcul correspond à la séquence des statistiques dérivées. On sait que dans cet ordre les valeurs sont dérivables.

Pour chaque statistique de la séquence des statistiques dérivées et selon cette ordre :

- Vérifier que toutes les statistiques de base ont été introduites (validées).
- Calcul des statistiques dérivables à partir des constantes sous contrainte structurelle. Voir le point 3.1.2. b) Les constantes et les équations sous contrainte structurelle.
- Calculer les valeurs des statistiques dérivables.
- La validité par rapport aux contraintes, énumérées au point *3.1.2. c) Les contraintes structurelles* est vérifiée.

4.4.5. Algorithme de calcul et de validation des statistiques de l'instant T_i

Le calcul et la validation des statistiques de l'instant T_i ne pose pas de problème particulier (c'est pourquoi il n'a pas été abordé dans l'analyse). Le calcul du modèle statistique des valeurs à l'instant T_i demande la connaissance du modèle statistique des valeurs de base à l'instant T_o. Le calcul se réalise comme suit :

- Les statistiques de base de l'instant T_i sont calculées avec la formule $s_i = s_o + d(s_o) \times (T_i - T_o) / T$. Chaque statistique obtenue est validée par rapport aux contraintes structurelles.

- Le calcul des statistiques dérivées se réalise de la même façon que dans l'algorithme précédent.

4.4.6. Algorithme de transformation

L'algorithme de transformation concerne la destruction et la création des statistiques, lors de la transformation d'un schéma.

- Déclasser les anciennes statistiques de base, de la même manière que dans l'algorithme abstrait de construction.

- Classer les nouvelles statistiques de base, de la même manière que dans l'algorithme abstrait de construction. Les nouvelles statistiques de base sont celles obtenues avec des anciennes statistiques de base, elles sont repérées soit avec les équations de transformation, soit (pour les équations de transformation à plusieurs arguments) par les équations de dérivation où tous les arguments sont des statistiques de base.

4.4.7. Aperçu général de la gestion abstraite des statistiques

La figure ci-dessous montre différents aspects de la gestion des statistiques. D'une part, il y a les composants abstraits de gestion dans la première colonne. D'autres part il y a les algorithmes abstrait qui opèrent sur le modèle statistique, lequel fait partie des informations disponibles sur les bases de données dans la base de spécifications.

Une première remarque : la construction d'un modèle statistique des équations n'y paraît pas. Une telle construction est implicite, elle se réalise au fur et à mesure de la construction du modèle statistique de classement, directement de la base de données à partir des informations de la base de spécifications.

Une deuxième remarque porte sur la notion de non-redondance du modèle statistique. Elle est implicite étant donné que grâce à l'algorithme décrit plus haut, les modèles sont toujours non-redondants.

Suivons maintenant le scénario que la gestion abstraite nous permet de réaliser :

- Construction d'un modèle statistique de classement
- Le modèle statistique des valeurs de base est établi. Les valeurs ne vérifiant pas les contraintes structurelles sont refusées.
- Sur demande le calcul des statistiques s'effectue. Le modèle statistique est refusé s'il y a violation de contrainte structurelle et nous recommençons le point précédent : introduire des statistiques de base

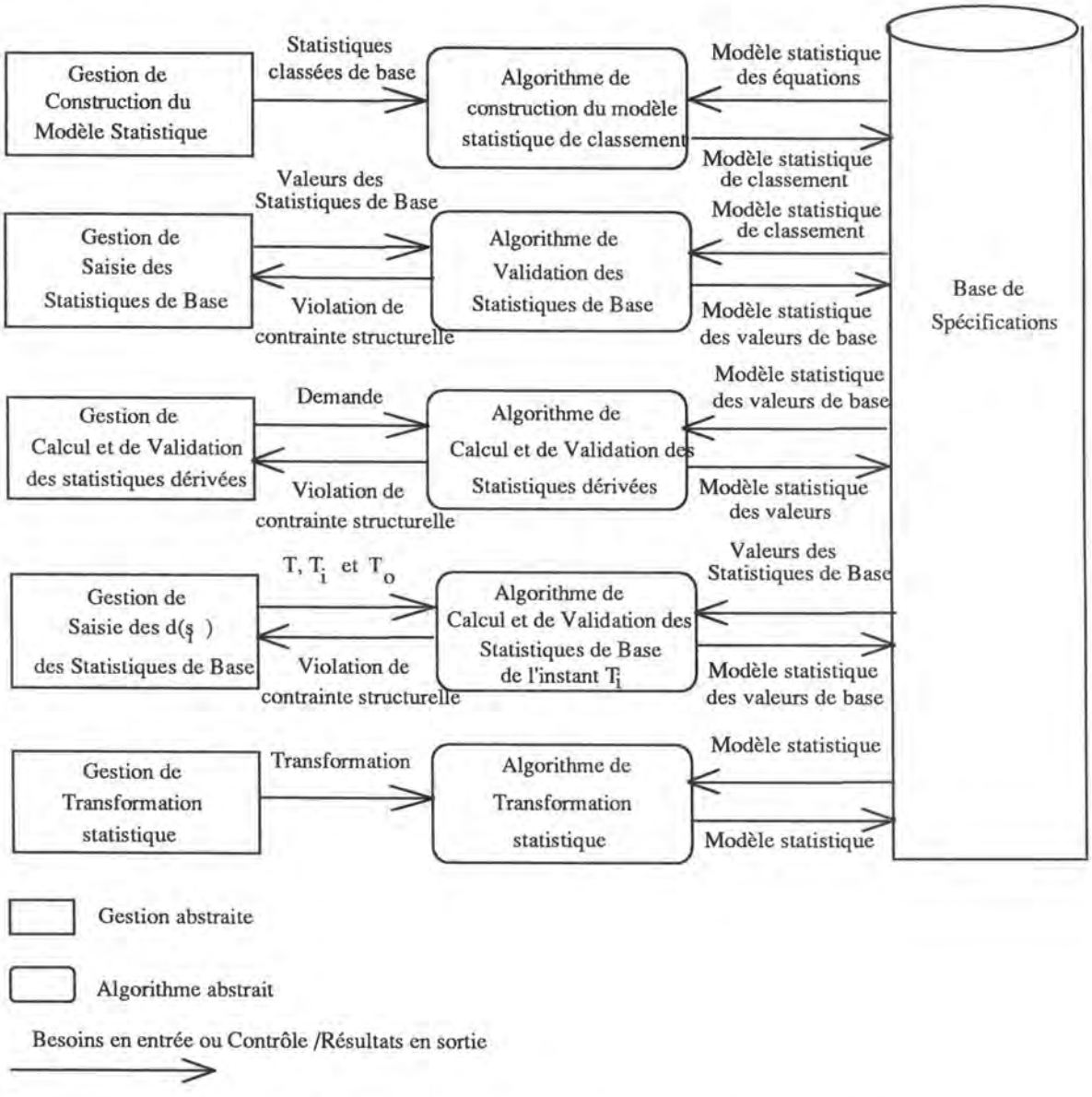


Figure 4.17. : *Aperçu général de la gestion des statistiques*

4.5. Commentaires sur les algorithmes

Optimisation

La gestion doit pouvoir offrir à l'utilisateur les statistiques de base à l'origine d'une statistique dérivée dont la valeur est non-valide. Le graphe des dépendances permet facilement de sélectionner le sous-graphe composé des ascendants de la statistique dérivée dont la valeur est non-valide. Toutefois, les ascendants d'une statistique dérivée peuvent également être retrouvés avec la structure abstraite (plus simple) du modèle statistique; et le graphe des dépendances en tant que tel n'est donc pas nécessaire pour la gestion.

Amélioration

Déclasser une statistique de base demande de refaire le classement de la moitié, en moyenne, des statistiques de base. Une optimisation possible de l'algorithme abstrait pour le déclassement d'une statistique de base, consiste à utiliser un graphe des dépendances construit lors de la construction du modèle statistique de classement. Nous n'avons pas envisagé cette optimisation dans le mémoire puisqu'elle n'est pas indispensable pour la gestion.

Alternative

Nous avons proposé la transformation du modèle statistique comme gestion des statistiques entre différentes bases de données dans les points *4.3.7. Transformation du modèle statistique* et *4.4.6. Algorithme de transformation* . Nous pensons que la transformation du modèle statistique n'est pas toujours utile. Une alternative s'offre quand on gère les différents schémas du processus de conception (ce qui serait une évolution fort probables pour les ateliers logiciels de conception) et que l'apport des statistiques se fait pour l'un deux pour être calculées ensuite pour les autres. Le calcul correspond alors à l'application des équations de transformation.

Chapitre 5

Application à l'environnement TRAMIS

Ce chapitre nous présente l'analyse de deux propositions d'intégrations de la gestion des statistiques mis en évidence dans le chapitre précédent. La structure de données de TRAMIS, la base de spécifications, est étudiée pour l'intégration d'une structure de gestion statistique.

5. Application à l'environnement TRAMIS

5.1. Introduction

Maintenant que toute la gestion des statistiques est énoncée, passons aux structures de données des statistiques.

Le choix entre les deux possibilités d'intégration dans la base de spécification de TRAMIS : les statistiques en tant qu'attributs des objets existants ou en tant qu'objets statistiques indépendants, n'est pas facile à faire. Une analyse plus approfondie est nécessaire pour pouvoir évaluer les différents avantages.

La structure du modèle statistique de classement établie par l'algorithme doit répondre au besoin de la construction, au besoin du calcul et doit s'intégrer dans la structure de données de TRAMIS au mieux. Le type de cette structure demande donc un choix judicieux, c'est pourquoi il est mis en évidence dans ce chapitre.

5.2. La base de spécification de TRAMIS.

Dans le cadre du mémoire, la conception de base de données est à la fois sujet de matière ainsi que l'outil de travail. L'outil de travail est la base de spécification de TRAMIS.

TRAMIS est un atelier logiciel destiné à aider les concepteurs d'un système d'information à concevoir des bases de données. Ces bases de données en projets sont mémorisées (et manipulées) dans une base de données des spécifications. Cette base de donnée des spécifications a été conçu indépendamment des traitements qui l'exploitent (les programmes de TRAMIS). Elle permet donc de stocker toutes les informations d'un schéma. Nous allons commenter la partie de la base de spécifications de TRAMIS qui nous est utile pour la gestion des statistiques.

La description qui suit est basée sur le rapport technique [DEC-MAR,92] et couvre les concepts de la partie du méta-schéma en question.

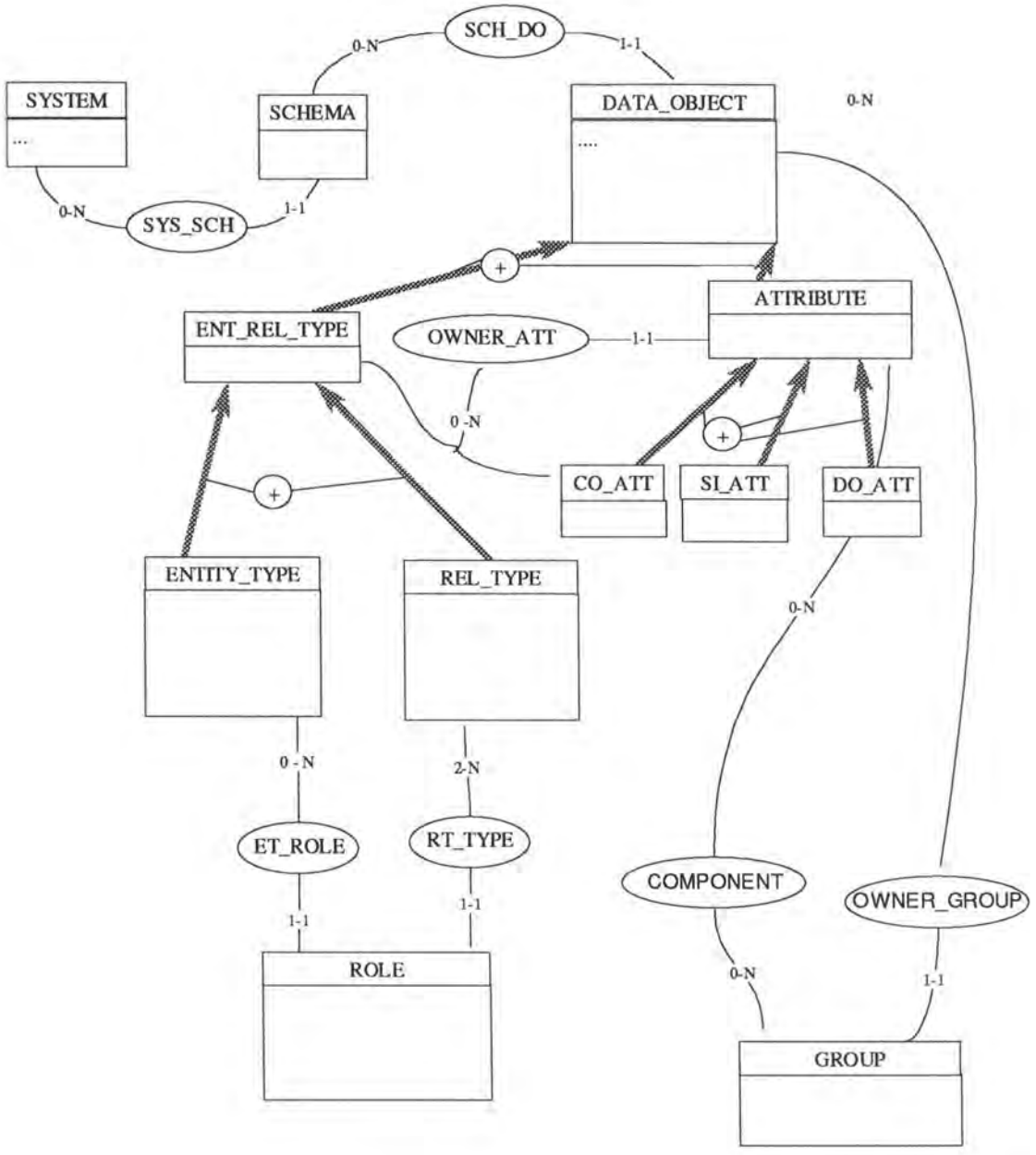


***Figure 5.1.** : Le méta-schéma de la base de spécification*

La base de donnée contient la description les schémas d'un ou plusieurs systèmes d'informations (SYSTEM). Un SYSTEM possède une ou plusieurs descriptions de données (SCHEMA). Une description de données prend la forme d'un SCHEMA attaché à un SYSTEM (via SYS_SCH).

L'élément de base d'un schéma est la donnée (DATA_OBJECT). Tout DATA_OBJECT appartient à un schéma (via SCH_DO). Le DATA_OBJECT, est en fait la généralisation de type d'entité (ENTITY_TYPE), type d'association (REL_TYPE)

et attribut (ATTRIBUTE). Ces trois types d'objets spécifiques sont donc regroupés en un objet générique, dont ils forment une partition.

Les types d'entités ENTITY_TYPE et REL_TYPE sont généralisés en un type d'entité générique ENT_REL_TYPE. Un ENT_REL_TYPE possède zéro ou plusieurs attributs (via OWNER_ATT).

Toute propriété caractérisant un type d'entité ou un type d'association est représentée par un attribut (ATTRIBUTE) qui lui est associé via OWNER_ATT.

Il existe trois types d'attributs : les attributs décomposables (CO_ATT), les attributs élémentaires (SI_ATT) et les attributs associés à un domaine particulier (DO_ATT). Ces trois objets sont la spécialisation de l'objet générique ATTRIBUTE dont ils forment une partition.

Un groupe (GROUP) est un ensemble d'attributs (ATTRIBUTE) et/ou de rôles (ROLE) et/ou de groupes (GROUP). Chacun de ces attributs, rôles ou groupes peuvent appartenir à plusieurs groupes. Cette notion est modélisé par le type d'association COMPONENT qui relie un groupe à ses composants. Un groupe est associé à un DATA_OBJECT via le type d'association OWNER_GROUP.

5.3. Représentation des statistiques

5.3.1. Les statistiques en attributs dans la base de spécifications

a) Partie statistique de la base de spécifications

Les statistiques sont les attributs des objets, du méta-schéma de la base de spécification de TRAMIS. (L'objet Domaine n'est pas pris en compte par la gestion actuelle, c'est pourquoi nous n'allons pas considérer les statistiques de l'objet domaine par la suite.)

- Les statistiques pour l'objet SCHEMA

T_o, T_i, T

- Les statistiques pour l'objet ENT_REL_TYPE

N_{ER}

- Les statistiques pour l'objet ATTRIBUT

N_A, λ_A, $\mu_{A/ER}$, $\mu_{ER/A}$, $\Pi o_{A/ER}$, $\mu'_{A/ER}$, λ'_A

- Les statistiques pour l'objet ET_ROLE

μ_{ri}, Πo_{ri}, μ'_{ri}

- Les statistiques pour l'objet GROUPE

N_G , μ_G

b) Le type de la structure du modèle statistique

Le mot séquence a été intentionnellement choisi dans le chapitre précédent, au lieu de liste ou tableau puisque cela implique un choix d'implémentation qui n'est pas évident à faire.

Un élément de la séquence doit correspondre avec une statistique du schéma. Envisageons le cas ou les statistiques feront partie du méta-schéma en tant qu'attributs des objets. Une statistique d'un schéma est obtenue par un pointeur vers l'objet et un numéro pour l'identifier parmi les autres statistiques de l'objet. L'espace mémoire d'un pointeur est de 4 bytes et pour le numéro 1 byte suffit. Donc pour un élément (statistique) il faut 5 bytes.

Pour un schéma dont on connaît les objets on peut calculer le nombre de statistiques. Dénombrons les statistiques (à partir du point 5.1.1. a) *Partie statistique de la base de spécifications*) par objet :

OBJET	nombre de statistiques
ENT_REL_TYPE	1
ATTRIBUT	7
GROUPE	2
ET_ROLE	3
SCHEMA	3

***Table 5.1.** : Nombre de statistiques par objets*

Prenons en moyenne 3 statistiques par objets et un schéma de 10 000 objets. Cela fait en tout 30 000 statistiques.

Envisageons les deux cas d'implémentations, liste ou tableau :

- Une liste de pointeurs comme implémentation des séquences

L'exemple du chapitre précédent, Figure 4.11. : Exemple de structure du modèle statistique, prendra alors la forme suivante :

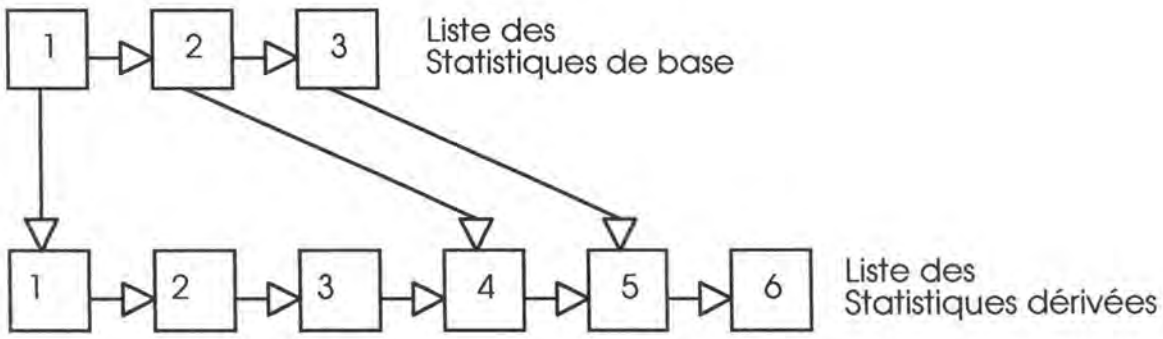


***Figure 5.2.** : La séquence comme liste de pointeurs*

Le calcul que l'on peut faire, avec 4 bytes par pointeur pour le chaînage des listes et un nombre 10 000 de statistiques de base et de 20 000 statistiques dérivées , pour évaluer le volume : 30 000 x (5 bytes + 4 bytes) + 10 000 x 4 = 310 000 bytes au total.

- Un tableau comme implémentation des séquences

L'exemple du chapitre précédent, *Figure 4.13. : Exemple de structure du modèle statistique*, prendra ici la forme montrée dans la Figure ci-dessous.

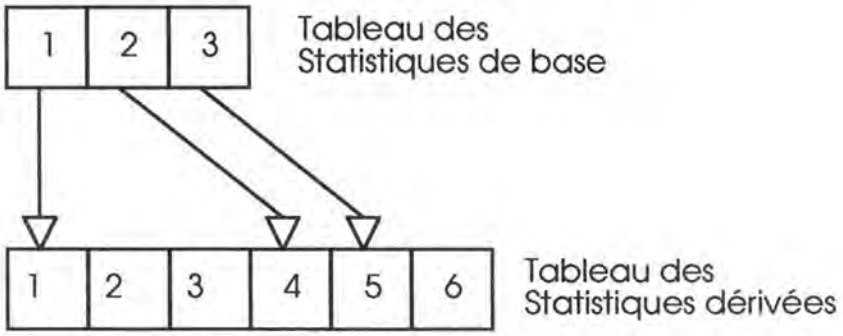


Figure 5.3. : *La séquence comme un tableau*

Le calcul que l'on peut faire ici, avec 4 bytes par pointeur pour les pointeurs des éléments de base (au nombre de 10 000), pour évaluer le volume 10 000 x (5 bytes + 4 bytes) + 20 000 x 5 bytes = 190 000 bytes au total.

- Des deux, le tableau prend le moins de place, mais le volume doit être réservé pour les statistiques. Ceci n'est peut-être pas trop encombrant si l'on considère que l'utilisateur se consacre entièrement aux statistiques. L'utilisation d'un tableau est aussi plus aisée.

c) La structure du modèle dans la base de spécifications

La structure du modèle statistique utilisée ci-dessus peut difficilement faire partie de la base de spécifications de TRAMIS si nous voulons l'utiliser comme tel. Nous allons dans ce point chercher une forme adéquate pour intégrer le modèle dans la base de spécifications. On doit donc pouvoir passer facilement et de manière univoque d'une spécification (représentation) à l'autre.

Résumons ce que nous avons déjà énoncé sur le modèle statistique. D'une part il y a le concept de classe qui nous renseigne si la statistique est de base, dérivée ou indéterminée. D'autre part il y a la séquence des statistiques de base et la séquence des statistiques dérivées.

- La classe peut être représentée par une variable.

- Les séquences peuvent être construites à partir d'un numéro d'ordre, puisque les séquences sont chronologiques. Les liens de la séquence des statistiques de base vers la séquence des statistiques dérivées peuvent être introduits par des sauts dans le numérotage. Ainsi le pointeur de la troisième statistique de base pointe vers la première statistique dérivée après le deuxième saut.

Pour le même exemple à nouveau :

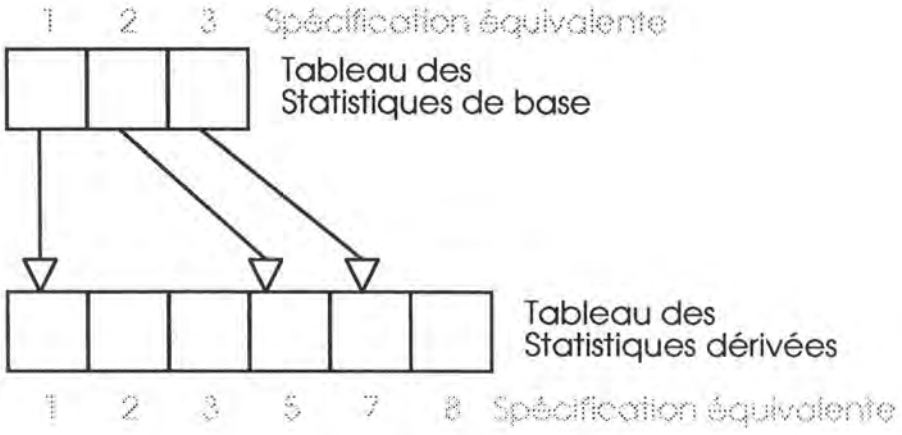


Figure 5.4. : *Une spécification équivalente du modèle statistique dans la base de spécifications*

Toutefois la classe et les listes peuvent être intégrées dans la base de spécifications avec un seul entier. Employons les entiers positifs pour numéroter les statistiques de base, les entiers négatifs pour les statistiques dérivées et la valeur nulle pour indiquer les statistiques indéterminées. Donc un seul entier, appelons le M (de modèle), suffit pour chaque statistique.

Modifions encore un petit peu la structure utilisée par le programme de construction.

- N'employons plus deux tableaux, un seul.
- Au lieu d'utiliser un pointeur pour les liens entre statistique de base et statistiques dérivées, introduisons des éléments vides qui sont chaînés entre eux.

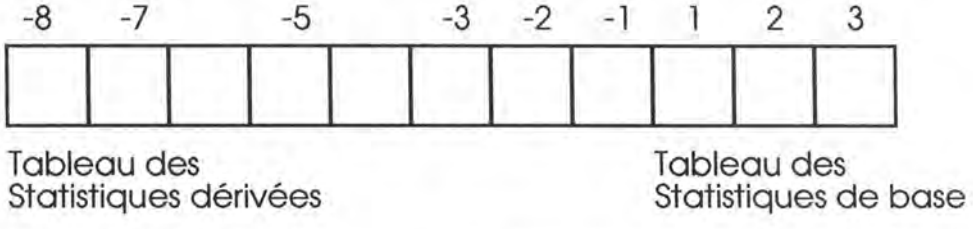


Figure 5.5. : *La spécification équivalente adoptée du modèle statistique dans la base de spécifications*

La construction de la structure d'un modèle à partir de cette entier demande un parcours de toutes les statistiques du schéma. Les éléments (pointeur de l'objet et numéro de statistique) sont facilement remplies dans le tableau à la position indiquée par l'entier, pour les valeurs non nulles.

d) Les opérations sur la structure du modèle statistique

Nous allons approfondir comment réaliser les deux opérations élémentaires de l'algorithme et vérifier si l'utilisation de la structure de tableau comme structure du modèle statistique est aisée ou non. Pour faciliter la compréhension les opérations sur le tableau sont commentées encore sur l'exemple du chapitre précédent, *Figure 4.11. : Exemple de structure du modèle statistique*, sous sa nouvelle structure.

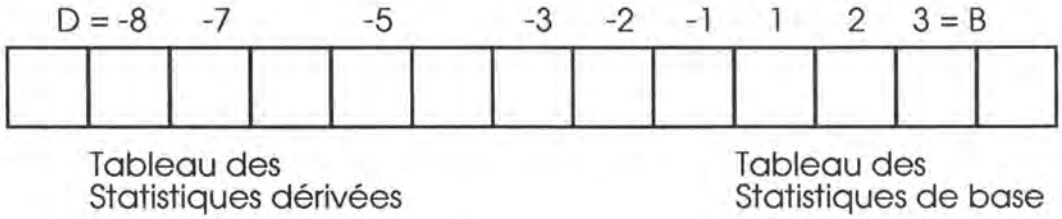


Figure 5.6. : *Les opérations sur la structure du modèle statistique*

Les variables D et B sont respectivement le numéro de la dernière statistique dérivée et de la dernière statistique de base.

1. Classer une statistique comme statistique de base
   - Rajouter une statistique de base.
   - Rajouter les statistiques dérivables.

   Dans le tableau cela ne pose pas de problèmes :

   ```
   B = B + 1
   copier un pointeur sur l'objet et le numéro de la statistique dans l'élément
   D = D +1
   pour chaque statistique dérivée
       D = D +1
       copier un pointeur sur l'objet et le numéro de la statistique dans
       l'élément
   ```

   Dans le méta-schéma ces opérations reviennent à faire la mise à jour de la valeur B ou -D pour l'entier M.

2. Classer une statistique de base comme statistique indéterminée.
   - Enlever une statistique de base.
   - Enlever les statistiques qui ne sont plus dérivables.

   Dans un tableau cela ne pose pas de problèmes particuliers de décaler tous les éléments après un élément d'une place en arrière :

   ```
   Pour chaque statistique S du tableau > M décaler en arrière.
   Trouver le (M-1)ième vide V dans le tableau
   ```

$|D = V + 1$

La réalisation du décalage peut se faire par copie de zone mémoire en une instruction pour ne pas compromettre la performance du point de vue du temps de traitement.

Dans le méta-schéma ces opérations reviennent à remettre l'entier M à la valeur nulle pour ces statistiques.

e) Le type statistique

Sauf pour les statistiques de l'objet SCHEMA qui sont de type réel, le type des statistiques est composé d'une valeur initiale $\mathbf{s_0}$, d'une valeur courante $\mathbf{s_i}$, un incrément par période $\mathbf{d(s_i)}$ et la variable **M** représentant le modèle.

TYPE STATISTIQUE

s_0 : réel	s_i : réel	$d(s_i)$: réel	M : entier

Figure 5.7. : *Le type statistique*

5.3.2. Un type objet statistique dans la base de spécifications

a) Partie statistique de la base de spécifications

Les statistiques forment un type d'objet et sont reliées avec des associations aux objets auxquels elles appartiennent. Les objets statistiques sont créés au fur et à mesure que l'utilisateur en a besoin. C'est à dire que qu'il n'y a que les statistiques de base et dérivées qui sont créées.

Les statistiques sont gérées plus indépendamment des type objets. Cela allège les types objets, en particulier le type d'objet ATTRIBUT.

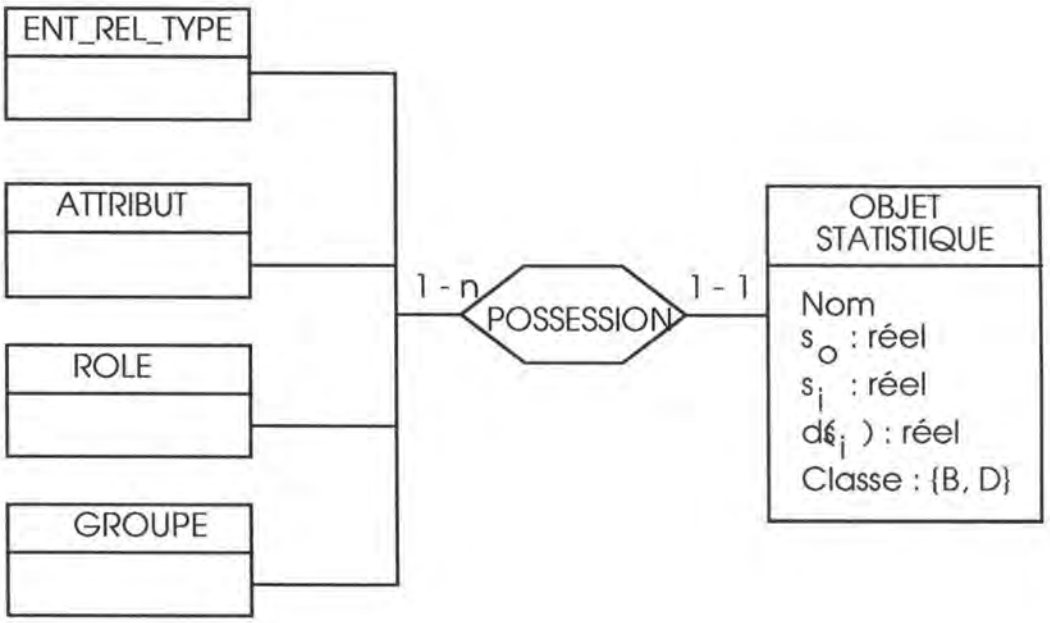


Figure 5.8. : *Le type objet statistique*

b) Le type de la structure du modèle statistique

La structure du modèle statistique dans ce méta-schéma prend la forme d'une association triée et accompagnée d'un attribut classe. Pour l'exemple déjà rencontré, *Figure 4.13. : Exemple de structure du modèle statistique*, les statistiques sont triées selon la suite d'entiers.

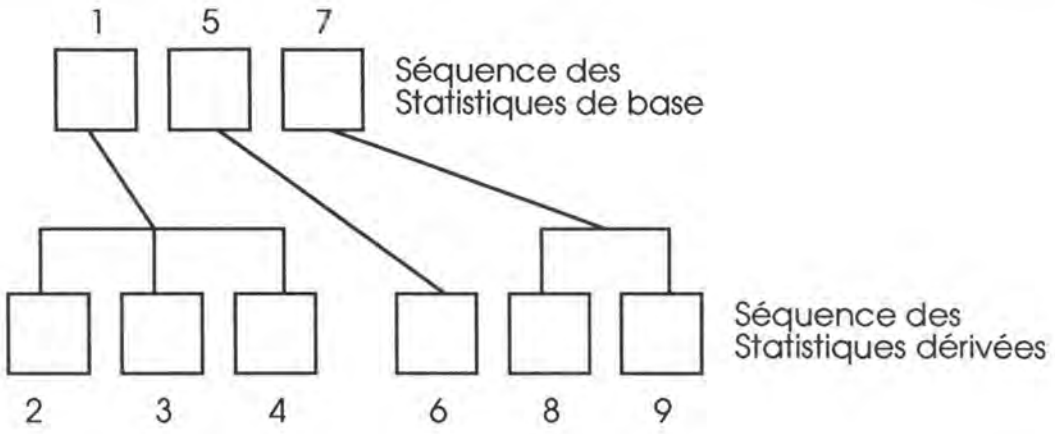


***Figure 5.9.** : La séquence triée sur l'objet*

Le type d'association ORDRE permet que les occurrences de l'OBJET STATISTIQUE soient accessible dans l'ordre de la structure du modèle statistique.

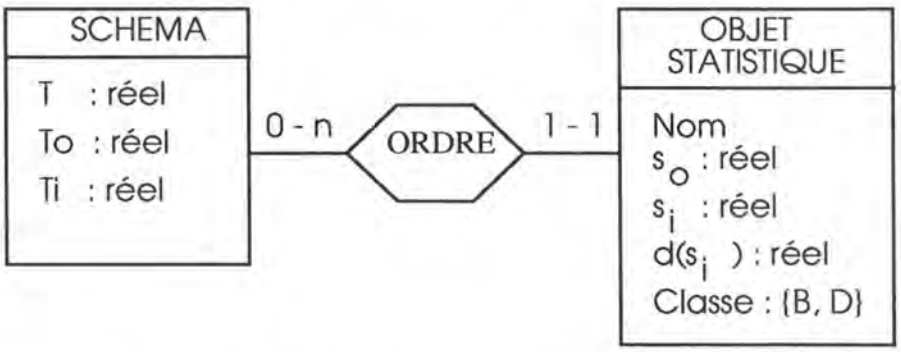


***Figure 5.10.** : Le type de la structure du modèle statistique*

c) Les opérations sur la structure du modèle statistique

Nous allons approfondir la réalisation les deux opérations élémentaires de l'algorithme et vérifier si l'utilisation de la structure du modèle statistique est aisée ou non.

1. Classer une statistique comme statistique de base
   - Rajouter une statistique de base
   - Rajouter les statistiques dérivables

Ne pose aucun problème

2. Classer une statistique de base comme statistique indéterminée.
   - Enlever une statistique de base
   - Enlever les statistiques qui ne sont plus dérivables

Destruction de l'objet statistique de base.
Destructions des objets statistiques des statistiques dérivées.
Créations des objets statistiques dérivables pour chaque statistique de base qui suit celle détruite.

5.3.3. Comparaison

Le choix entre les deux possibilités d'intégration ne peut pas encore être fait, parce que nous ne n'avons pas encore analysé le comportement de l'utilisateur, lequel est surtout intéressé d'arriver à une solution (modèle statistique des valeurs valides) facilement et sans trop de temps d'attente. Nous pouvons néanmoins déjà énumérer certains données :

1. **Statistiques en tant qu'attributs des objets**
- Les statistiques indéterminées prennent de l'espace mémoire
- Le modèle est représenté par un entier et un tableau indépendant
- Accès direct au statistiques à partir des objets
- Copie (décalage) de zone mémoire

2. **Statistiques dans un type d'objet statistique**
- Les associations entre l'objet statistique et les objets auxquelles il appartient prennent de l'espace mémoire
- Le modèle statistique est représenté par une association triée et un attribut classe
- Accès aux statistiques via des associations à partir des objets demande du traitement de temps
- Destruction des objets statistiques dérivées (et leurs liens !)

3. **L'algorithme abstrait**
- Classer une statistique de base comme statistique indéterminée demande de reclasser en moyenne la moitié des statistiques.

5.4. Architecture et dialogues

Cette section aborde le problème de l'intégration de la gestion des statistiques dans l'interface de TRAMIS, et correspond à l'analyse de différents scénari de gestions que l'utilisateur peut suivre. Cette section peut éclairer les parties de la gestion qui sont restées obscures pour le lecteur.

5.4.1. Aperçu général de la gestion des statistiques

Commentons d'abord les différentes composantes d'interface que la gestion des statistiques nécessite. L'utilisation de celles-ci suivra dans les différentes fonctions que la gestion implique.

Les composantes de gestion abstraites que nous avons commenté plus haut, se concrétise ici par des fonctions et les algorithmes abstraits par des processus de programmes.

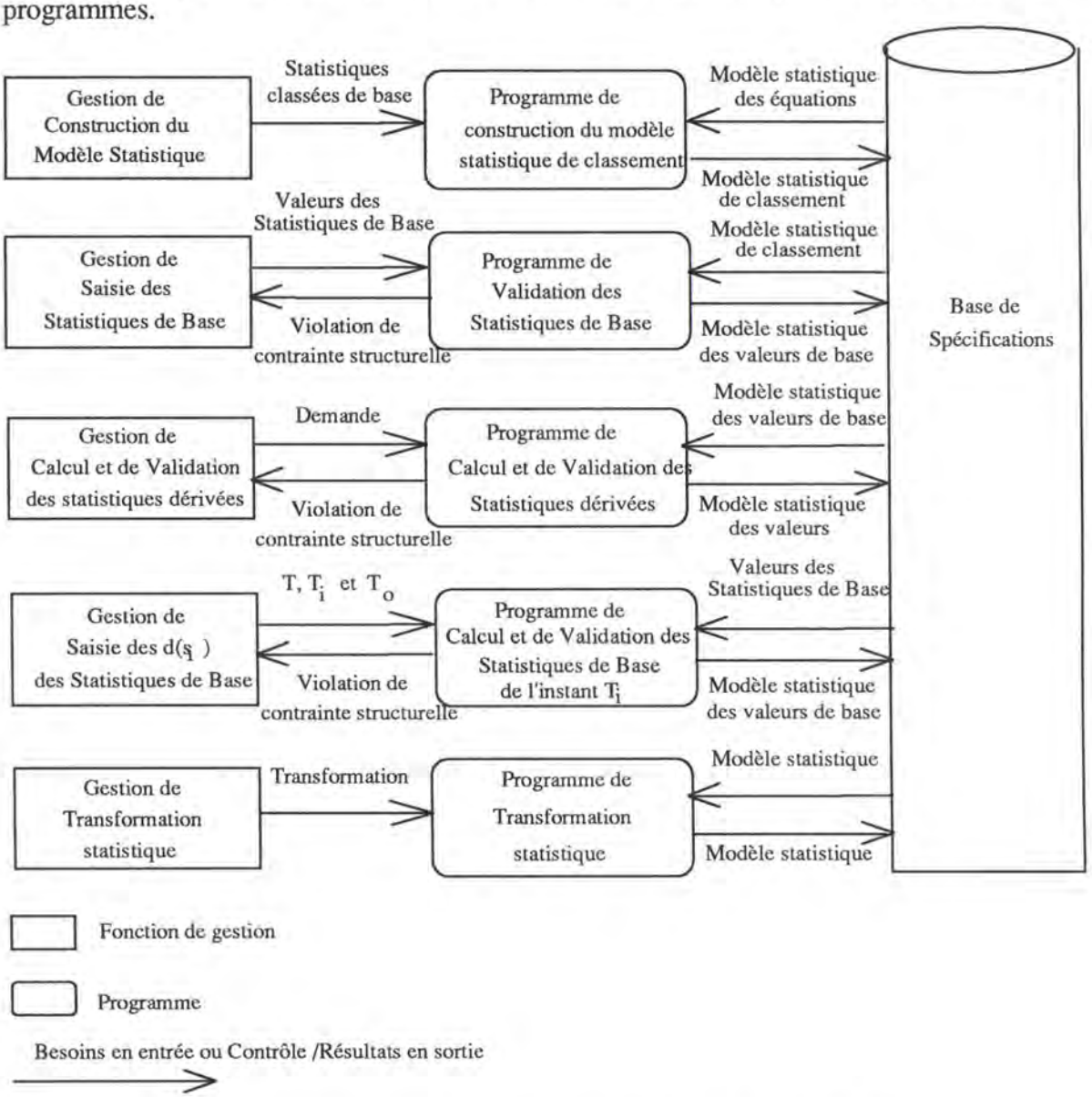


Figure 5.11. : *Aperçu général des fonctions de gestion des statistiques*

Citons les objets d'interface que les différentes fonctions nécessitent :

- L'objet d'interface Fenêtre est utilisé dans la fonction de gestion de construction du modèle.
- L'objet d'interface Boîte de saisie est utilisé dans la fonction de gestion de saisie des statistiques de base.

- L'objet d'interface Fenêtre est utilisé dans la fonction de gestion de calcul et de validation des statistiques dérivées, pour afficher les statistiques de base à partir desquelles une valeur non-valide est dérivée.

- L'objet d'interface Boîte de saisie est utilisé dans la fonction de gestion de saisie des incrément par période $d(s_i)$.

- La fonction de transformation statistiques est transparente pour l'utilisateur. Elle intervient implicitement à la transformation d'un schéma (ou d'une partie d'un schéma).

5.4.2. Scénario de construction

La construction du modèle statistique de classement se fait interactivement par l'utilisateur à partir de toutes les statistiques de la base de données. Les statistiques qui ne sont pas (encore) classées sont les statistiques indéterminées.

La fonction statistique **Construction modèle** au niveau du schéma comporte : **Une fenêtre résidu**, contenant toutes les statistiques indéterminées et **une fenêtre modèle-statistique**, contenant toutes les statistiques de base ou dérivées du schéma. Les statistiques dérivées de la fenêtre modèle-statistique ne peuvent pas être sélectionnées (elles sont inactives, par exemple en gris clair). L'exemple de construction,*Table 4.2. : Modèle statistique de classement de l'exemple*, jusqu'à la statistique N_{G1} (16) .

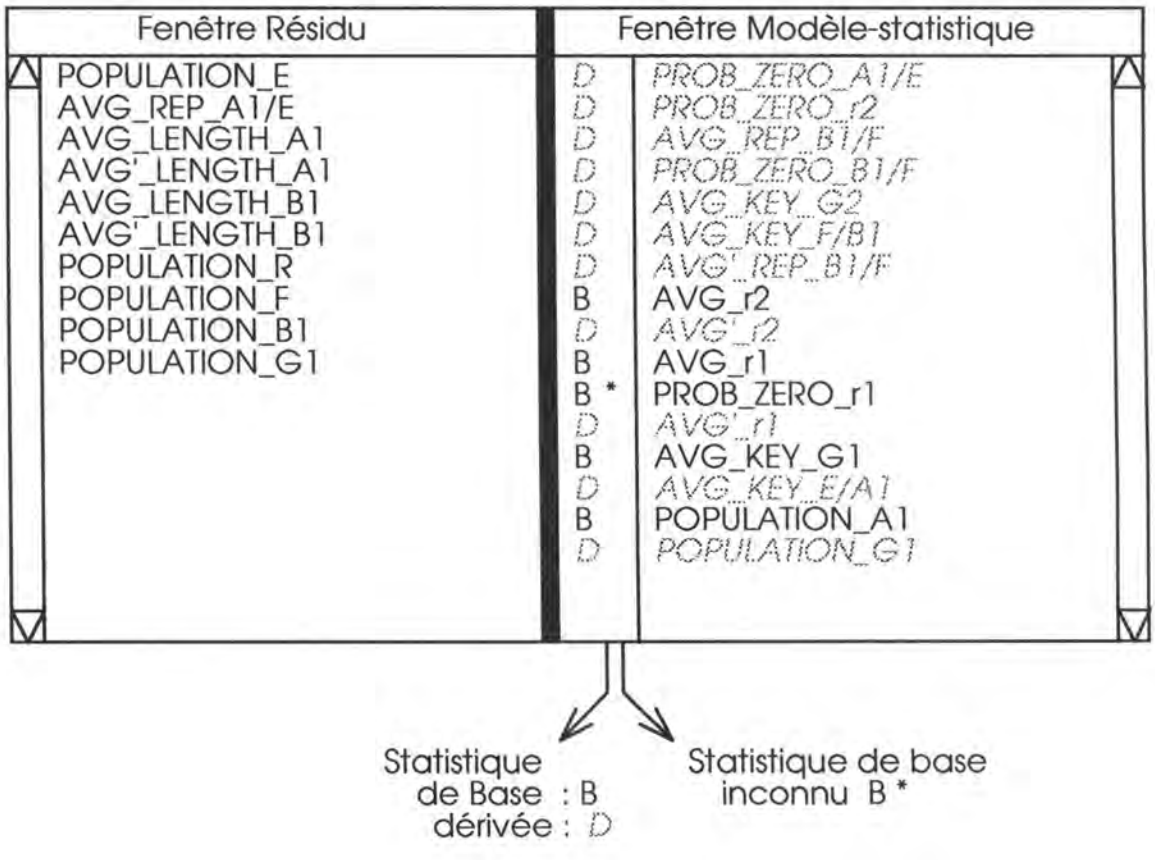


***Figure 5.12. :** Fenêtres de construction du modèle*

Avec comme fonctions :

1. reset all : Remettre toutes les statistiques dans la fenêtre résidu.
2. set base + élément sélectionné de la fenêtre résidu :
   Mettre la statistique et les statistiques dérivables dans la fenêtre modèle-statistique par une Boîte de confirmation.
3. reset base + élément sélectionné de la fenêtre modèle-statistique :
   Remettre la statistique et les statistiques dérivables dans la fenêtre résidu.
4. saisie + élément sélectionné de la fenêtre modèle-statistique :
   Accède à la boîte de saisie.
5. cancel : Ferme les deux fenêtres sans mise à jour.
6. ok : Ferme les deux fenêtres avec mise à jour.

5.4.3. Saisie et validation des statistiques

a) Les statistiques statiques

La saisie des statistiques concerne les statistiques de base. Elle est réalisée après ou pendant la construction du modèle statistique et respecte la structure élaborée. La saisie est validée par rapport aux contraintes structurelles.

Pour ne pas avoir de Boîte de saisie différente pour chaque statistique, nous proposons de prendre une Boîte de dialogue par objet. De la table, Table 5.1. : Nombre de statistiques par objets, nous pouvons conclure que l'utilisateur sera confronté avec 5 boîtes de dialogues différentes.

Un ***bouton statistique*** au niveau de la saisie des objets d'un schéma de TRAMIS donne accès à la boîte de dialogue pour les fonctions : saisie, modifie et consultation des statistiques. (Pour les statistiques appartenant à un même objet la même boîte de dialogue est activée.)

Un autre bouton statistique dans la boîte de confirmations lors d'un classement d'une statistique comme statistique de base, donne accès à la Boîte de saisie.

Un troisième bouton statistique est offert dans la fonction construction modèle, pour les statistiques de base encore marquées inconnues.

b) Les statistiques d'évolution

La saisie de l'incrément par période d se fait pour chaque statistique de base dans sa Boîte de saisie. Les accès sont les mêmes que pour la saisie des valeurs des statistiques de base.

Les saisies des instants T_0 et T_i; et de la période T se font au niveau du schéma dans la fonction de gestion de saisie des $d(s_i)$ des statistiques de base. La Demande de calcul passe par un Boîte de confirmation, laquelle possède un bouton statistique par lequel on accède à la Boîte de saisie. Ainsi différents calculs sont facilement exécutables en variant uniquement ces dernières valeurs.

Chapitre 6

Conclusion

Dans ce chapitre nous commentons la complétude et les perspectives de ce travail.

6. Conclusion

6.1. Apport de ce travail

L'apport de ce travail réside principalement dans l'analyse des problèmes qui surviennent lors de la gestion de la description statistique d'une base de données existante ou en conception.

Nous nous sommes limités dès le départ aux statistiques statiques des objets (ou quantifications du contenu) d'une base de données. Les statistiques statiques peuvent être définies dans un schéma puisqu'elles dérivent de l'observation de la réalité que l'on modélise. C'est ce que nous avons fait dans le *Chapitre 3. Description statistique des données*. La représentation (des statistiques, des équations et des contraintes) et la manipulation (dans les transformations) y sont décrites d'une manière conceptuelle.

Le *Chapitre 4. Gestion des statistiques* constitue la partie la plus importante du travail. Son originalité réside dans l'analyse des problèmes que pose une gestion souple des statistiques. Des solutions aux différents problèmes sont proposées, et une de ces solutions est élaborée par la suite (jusqu'à la structure de données dans le *Chapitre 5. Application à l'environnement TRAMIS).* La gestion souple des statistiques consiste à laisser à l'utilisateur le choix des statistiques dont les valeurs sont à introduire. L'analyse des problèmes rencontrés porte sur l'importance et les possibilités de gestion de la redondance d'information et sur la correction de l'ensemble des valeurs introduites.

Le *Chapitre 5. Application à l'environnement TRAMIS* analyse l'intégration des statistiques et de la gestion des statistiques dans TRAMIS.

6.2. Possibilités d'extension du travail

Nous ne pouvons pas encore choisir une solution parmi les propositions d'intégration dans l'atelier logiciel TRAMIS. L'optimisation qui peut être réalisée avec la construction d'un graphe des dépendances lors de la construction du modèle statistique de classement n'a pas été envisagée pour l'intégration. Nous avons essayé dans un premier temps de considérer des structures de données aussi simples que possible. L'analyse de la structure d'un graphe des dépendances adaptée à la gestion des statistiques peut toutefois apporter une alternative intéressante. Cette variante reste donc à analyser.

L'assistance de l'utilisateur lors de la correction d'une valeur dérivée non-valide consiste à offrir à l'utilisateur la liste des statistiques qui sont à l'origine de cette invalidité. L'utilisateur peut ainsi rectifier des statistiques à sa guise, mais il n'est pas assisté dans son nouveau choix : il n'a d'autre choix que la méthode essai-erreur.

Le problème qui survient lorsque l'ensemble des contraintes est contradictoire (aucun ensemble de valeurs n'existe), n'est pas développé. Nous nous rendons compte que nous ne pouvons pas laisser l'utilisateur chercher une solution qui n'existe pas.

L'analyse de ce problème dépasse toutefois le cadre du mémoire, étant néanmoins une précondition de la gestion des statistiques.

Le mémoire s'est limité à l'évolution temporelle linéaire. Pourtant d'autres lois d'évolution restent à analyser.

Au-delà de ce que nous avons étudié, nous pouvons observer plusieurs possibilités d'extension.

Nous avons voulu aider l'utilisateur à trouver une solution qui représente la description statistique de la base de données existante ou en conception. Une autre façon de voir est de vérifier la solution qu'il souhaiterait avoir, et de faire la mise au point des valeurs contradictoires. Ceci revient à permettre la redondance d'informations, et donc la possibilité de contradictions, dans les informations quantitatives que l'utilisateur apporte. Dans cette optique, la gestion des statistiques prend la forme d'une gestion d'aide active où l'utilisateur essaye de trouver la solution qui correspond le mieux à celle voulue.

Nous avons toujours considéré que la gestion des statistiques est un but en soi. Il l'est si l'on veut s'assurer que l'ensemble des statistiques est valide. Mais nous pensons que l'analyse de l'exploitation des statistiques peut aboutir à une gestion active pour effectuer de la conception de bases de données efficaces. --- Pour expliquer ce que nous voulons dire avec efficace, nous devons d'abord parler des statistiques dynamiques. Les statistiques dynamiques dérivent de l'utilisation des objets d'un schéma, et précisent le taux d'utilisation des données par les traitements dans l'hypothèse d'un profil d'exploitation défini. L'exploitation d'une base de données peut être décrite en termes d'accès aux objets ou de créations, de suppressions et de modifications d'objets par unité de temps.--- Une base de données efficace est une base de donnée dont la description statistique (statique et dynamique) répond le mieux au profil d'exploitation de la base de données que l'on conçoit. Les statistiques sont essentielles pour la conception de bases de données efficaces et une gestion active serait donc très utile.

Chapitre 7

Bibliographie

La bibliographie reprend les articles et les livres dans l'ordre d'apparition dans le texte.

7. Bibliographie

[BOD-PIGN,83] BODART, PIGNEUR, Conception assistée des applications informatiques, Tome 1, 1. Etude d'opportunité et analyse conceptuelle, Masson, 1983.

[HAI,86] HAINAUT J.L., Conception assistée des applications informatiques, Tome 2, 2. Conception de la base de données, Masson, 1986.

[HAI,89] HAINAUT J.L., Introduction à la théorie relationnelle des bases de données, (Notes provisoires), 1989.

[HAI,92] HAINAUT J.L., A temporal Statistical Model for Entity-Relationship Schemas, in Proc. 11th Int. Conf. on Entity-Relationship Approach, Karlsruhe, October 1992, LNCS 645, Springer Verlag 1992.

[HAI-CAD-DEC-MARb,92], HAINAUT, J.L., CADELLI, M., DECUYPER, B. MARCHAND, O., TRAMIS : a transformation-based database CASE tool, in Proc. 5th Int. Conf. on software Engineering and Applications, Toulouse, 7-11 December, 1992.

[CONCIS,90], CONCIS, TRAMIS, Un atelier de conception de bases de données. Version 1.0 . Manuel de référence, 1990. 37bis, rue du Prébuard; F-95100 Argenteuil.

[FICH], FICHEFET J., Théorie des graphes et Réseaux de Pétri, Institut d'Informatique.

[DEC-MAR,92], B. Decuyper, O. Marchand, Description du méta-schéma dans une perspective orienté objet de l'atelier logiciel "TRAMIS", Rapport technique, Institut d'Informatique, 1992.

[RASE,92], RASE B., Contribution à la réalisation d'un atelier de conception de bases de données, Institut d'Informatique, Mémoire de fin d'études, FUNDP Namur, Septembre 1992.

[HAI-CAD-DEC-MAR, 92], HAINAUT, J.L., CADELLI, M., DECUYPER, B. MARCHAND, O., Database CASE Tool Architecture : principles for flexible design strategies, in Proc. 4th Int. Conf. on Advanced Information System Engineering (CAiSE-92), Manchester, May 1992, Springer Verlag, LNCS, 1992.